AN INVESTIGATION OF SOME STATISTICAL METHODS OF EVALUATING THE RELIABILITY OF RESULTS OF LABORATORY TESTS OF PETROLEUM PRODUCTS

MARTIN J. DEUTCH

1965

AN INVESTIGATION OF SOME STATISTICAL METHODS OF EVALUATING THE RELIABILITY OF RESULTS OF LABORATORY TESTS OF PETROLEUM PRODUCTS

by

Martin J. Deutch

B.S., University of Notre Dame, 1948

Submitted to the Department of Chemical and Petroleum Engineering and the Faculty of the Graduate School of the University of Kansas in Partial Fulfillment of the Requirements for the Degree of Master of Science.

ACKNOWLEDGEMENTS

The author wishes to acknowledge his gratitude to Dr. Charles F. Weinaug, director of the Petroleum Management Program, and to Dean Wiley S. Mitchell, School of Business advisor for the Petroleum Management Program, for providing guidance, encouragement and inspirational goals throughout the course of his matriculation in the Graduate School.

Deepest expressions of appreciation are due Dr. Floyd W. Preston whose patient, persevering guidance was responsible for the successful completion of this investigation.

The author sincerely appreciates the sponsorship of the United States Navy which provided this priceless opportunity for his educational advancement.

Finally, the author wishes to acknowledge the many sacrifices made by his wife and children who temporarily relinquished many of their claims on the master of the family in order to gain a master of science.

TABLE OF CONTENTS

CHAPTER PAGE

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

INTRODUCTION

Quality surveillance of military petroleum products is the aggregate of measures to be applied to determine and maintain their quality. Quality surveillance programs are conducted in order that required petroleum products will be available in a condition suitable for immediate use. Their ultimate purpose is: (1) to insure that no life is ever lost or equipment damaged or destroyed through the use of contaminated or deteriorated petroleum products, and (2) to promote economy by minimizing the necessity of surveying or reclaiming any petroleum products because of contamination or deterioration.

The success of any such program is dependent upon several factors, not the least of which is the maintenance of the highest standards of reliability in the testing laboratories.

Importance of Laboratory Reliability

A chemical analysis has been compared to an elastic yardstick never giving the same result twice.[1] How is one to know then if laboratory tests are "right"? The fact is that any decision regarding a petroleum product based on laboratory test results is a decision under uncertainty.

Decisions under uncertainty always involve a risk of making the wrong decision. This is of particular concern when test results of a petroleum product border on acceptability limits. In order to properly evaluate the risk of misclassifying borderline material, it is important to know how much stretch or shrinkage to allow in reported test results. Common sense dictates that it is also important to reduce the risk by reducing the elasticity of the yardstick as much as is economically feasible. The economics of reliability control are probably most apparent when considering a commercial application for which the costs of reliability control and the costs of wrong decisions can be quite accurately computed. Consider a refinery laboratory where small deviations could be expensive ones. When mixing a blend, a small excess per sample unit of an expensive component could add up to many dollars in excess costs in a continuous process. J. T. Walter cites a report by one refinery of losses of one million dollars per year on a single operation due to quality give-away.[2] Conversely, a deficiency could cause rejection of a product by a customer and add the costs of reprocessing to the product.

In regard to military applications, consider the cost of delay in discharging a tanker's cargo or defueling a ship while additional samples are tested, if the first sample results indicated that the quality was suspect. As a more

sobering example, we might visualize a heavily loaded aircraft faltering on take-off and crashing because of loss of power due to vapor lock. This could result from mis-classification of unfit fuel based on unreliable test results.[3]

From these examples, the importance of the reliability of laboratory test results in any attempt to control quality should be evident.

A Current Effort to Control Military Laboratory Reliability

At the command level, the maintenance of the highest standards of reliability in testing laboratories is dependent upon the ability to detect apparent trends toward unreliability. In pursuit of this goal, a correlation testing program has been set up within a major military area command as a part of its quality surveillance program. Identical samples of aviation gasoline, motor gasoline, jet fuel, diesel fuel and lubricating oil are prepared and distributed tri-annually to each of ten participating laboratories. The results are summarized and the average value of all observations is determined for each test. Reproducability limits are then computed for those tests for which a method of determining reproducability limits is given in the applicable American Society for Testing and Materials (ASTM) Standard. Reproducability limits can be computed for about seventy five per cent of the tests. The test results falling outside of these limits are indicated by an asterisk. A Summary

of Laboratory Performance is prepared which tabulates by activity, the number of tests reported for which reproducability limits are computed and the per cent within reproducability limits. Each summary includes the tabular data for each of the two preceding series of tests as well as for the current series.

Purpose of the Thesis

The purpose of this thesis is to investigate some statistical methods of treating the data obtained through the military area command correlation testing program described above to extract more definitive information from them concerning the reliability of the participating laboratories' test results.

CHAPTER II

RELIABILITY

This chapter discusses types of measurement error and their effects, and defines the associated terminology as it will be used throughout the following chapters.

Also defined are repeatability and reproducability as used by the American Society for Testing and Materials.

CAUSES OF UNRELIABILITY

Scarborough points out that all measurements are subject to three kinds of error: systematic or constant errors, mistakes, and accidental errors.[4]

Systematic Errors

Systematic or constant errors are those which affect all measurements alike. In regard to laboratory test results, they could for example, be due to improperly calibrated equipment or due to consistent but incorrect operative techniques. Systematic errors are usually evident as a constant bias.

Mistakes

Mistakes or blunders are due to carelessness primarily in making or recording observations. The fact that they do not follow any law makes gross blunders recognizable as isolated data points. Minor mistakes, however, may be difficult to detect.

Accidental Errors

Accidental errors are those whose causes are unknown or undetermined. They are usually small and they are considered to follow the laws of chance. Consequently they are also referred to as chance errors or random errors.

The mathematical theory of errors deals with accidental errors only. That is to say, systematic errors and gross blunders are due to assignable causes and can therefore be optionally eliminated, controlled, or accepted. Accidental errors however, cannot be avoided and are bound to occur with a measurable probability.

COMPONENTS OF RELIABILITY

Reliability, precision, and accuracy have been defined in various ways. All are comparative or relative terms rather than absolute measures. Arbitrary scales for their measurement must be established based on predetermined standards.

Precision

Precision is a quality of a set of data that describes the degree of dispersion of the values. The lower the dispersion or scatter, the higher the precision. Single measurements cannot be considered to be "precise" or "not precise."

Accuracy

Accuracy is a quality of a single measurement or a series of measurements that expresses the degree to which the single measurement (or the average of the set of measurements) conforms to a predetermined "true" value. High accuracy implies close agreement to the predetermined standard.

Target Analogy

The relationship between precision and accuracy is best explained through use of the target analogy.

Figure 2-1 illustrates four groupings of twelve shots in a target. Target A illustrates a grouping which is precise but not accurate. The shots are in a tight cluster but considerably removed from the center of the target area. This is analogous to the accompanying frequency histogram of laboratory measurements in which the measurements are grouped close together but their average value is considerably removed from the true value of the property being measured.

Target B illustrates accuracy without precision. The shots cluster around the center of the target in a random fashion but are widely scattered. Likewise in the accompanying frequency histogram, measurements are relatively evenly distributed around the true value but are relatively widely dispersed.

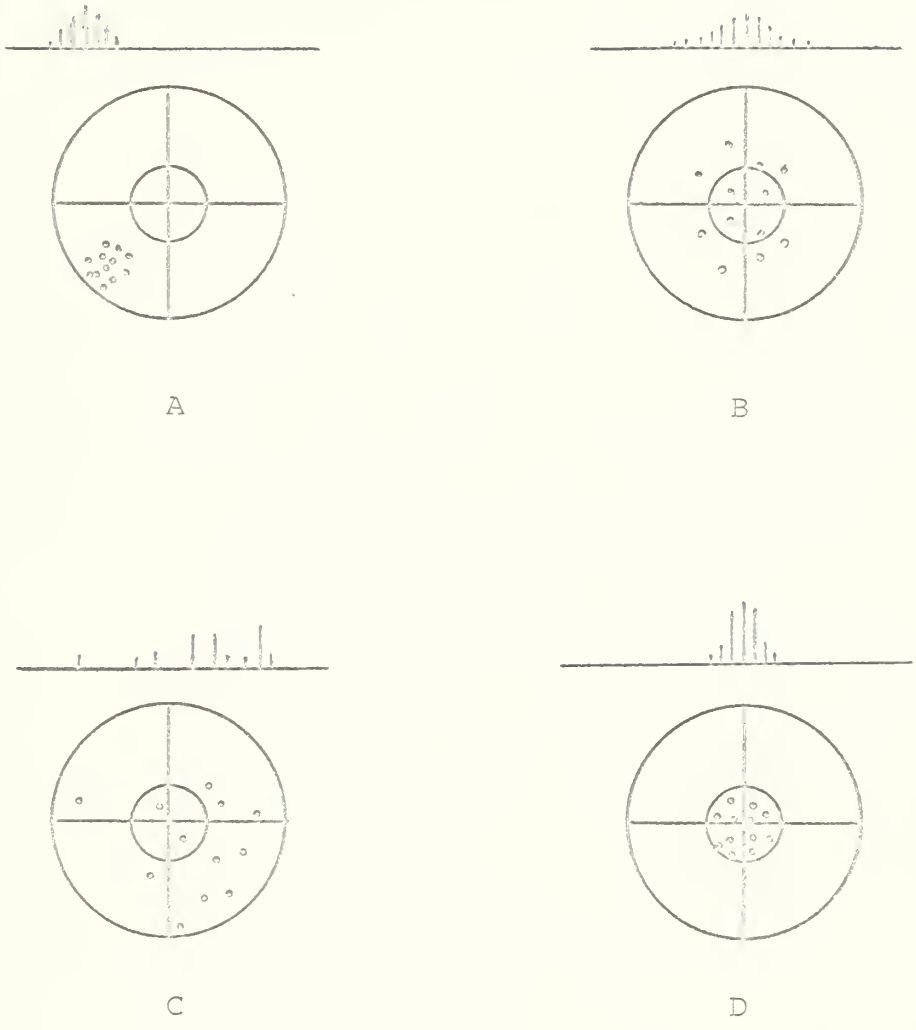

FIGURE 2-1

TARGET ANALOGY: PRECISION AND ACCURACY

Target C illustrates a dispersion of shots which is neither accurate nor precise. Again the shots are widely scattered and also do not form a uniformly dense pattern around the center of the target as they did in B.

Target D illustrates good marksmanship, that is marksmanship that shows high precision (tight clustering) and high accuracy (good centering).

Standards

The ASTM Standards on Petroleum Products and Lubricants provide convenient standards of precision in the form of Repeatability and Reproducibility amounts given with the description of the test method. Repeatability, is defined by them (ASTM) as the greatest difference between two single and independent results by a single operator in a given laboratory that can be considered acceptable at the ninety five per cent confidence level. Reproducibility, is defined as the greatest difference between a single test result obtained in one laboratory and a single test result obtained in another laboratory that can be considered acceptable at the ninety five per cent confidence level.

CHAPTER III

FUNDAMENTAL STATISTICAL MEASURES

Introduction

This chapter briefly discusses the fundamental statistical measures which are applied or considered in later chapters.

In the first part of the chapter the measures are defined. Methods of estimating population parameters from sample statistics are presented in the next section followed by a comparison of the relative efficiency of the various estimators. Finally a discussion is given of some of the advantages and disadvantages to be considered when choosing each statistic or estimator.

Frequent reference will be made to normal populations or distributions of values. The theory of the normal distribution stemmed from work done by Karl Gauss and, for this reason, the normal distribution is sometimes identified as the Gaussian distribution. The normal curve is defined mathematically as

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \text{exponential} - \frac{(x - \mu)^2}{2\sigma^2} \qquad (3\text{-}1)$$

in which μ is the mean value of the variable and σ is the standard deviation, both of which are described in this chapter.

In the context of equation (3-1) f(X) is known as a "probability density function." For any probability density function, f(X), the probability that a value of X lies in the interval $X_L \leq X \leq X_u$ is given by $P(X_L \leq X \leq X_u)$

$$P(X_L \leq X \leq X_u) = \int_{X_L}^{X_u} f(X)\, dX \qquad (3\text{-}2)$$

Thus, the probability that a value X lies between limits X_L and X_u is equal to the area under the probability density function f(X) between the two limits. This area is shown in Figure 3-1.

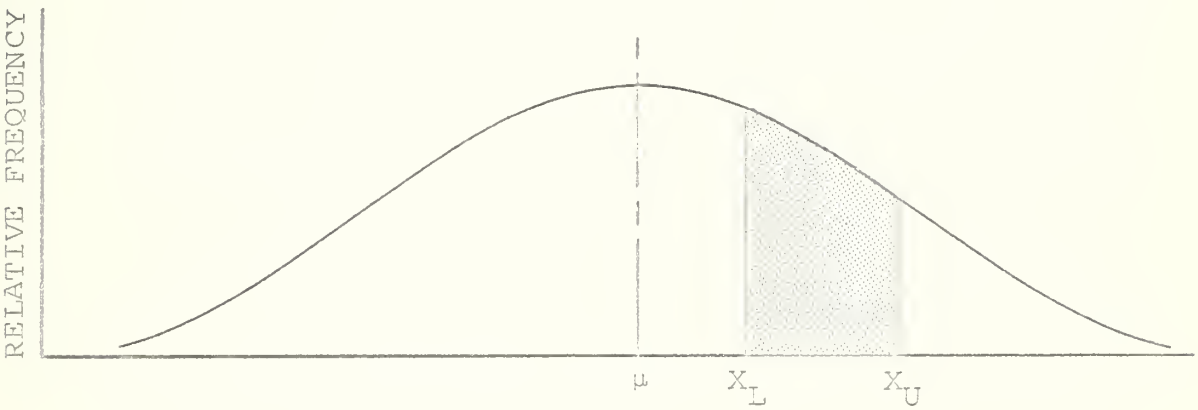

FIGURE 3-1

A PROBABILITY DENSITY FUNCTION SHOWING THE AREA EQUIVALENT TO THE PROBABILITY THAT X LIES BETWEEN X_L AND X_U

A "normal distribution" for a variable such as X signifies that the probability of X being between any two limits X_L and X_u is given by equation (3-2) if one uses equation (3-1) for the definition of f(X).

POPULATION PARAMETERS

Measures of Central Tendency

A universe or population is the totality of all pertinent observations that might be made in a given problem. If these observations are normally distributed, they will be symetrically dispersed around an "average" or central value. The central tendency of the population is of fundamental interest in any statistical analysis.

The ARITHMETIC MEAN or ARITHMETIC AVERAGE, μ, of a set of N values, X_i, is defined as the sum of the set of values, divided by the number of values in the set.

$$\text{POPULATION MEAN} = \mu = \frac{\sum_{i}^{N} X_i}{N} \qquad (3\text{-}3)$$

The arithmetic mean is the most commonly used measure of central tendency and is the value generally intended when the term "average" or "mean" is mentioned.

The MEDIAN is the middle value of a set of numbers arranged in ascending or descending order according to value. For an even number of data points, it is the arithmetic average of the two middle values.

50% of values < M < 50% of values

The MIDRANGE is a point halfway between the largest and smallest observations. It is computed as the average of

the first and last values of a set, ordered according to value

$$\text{MIDRANGE} = \frac{X_1 + X_N}{2} \quad \text{Where } X_1 < X_2 <. \; . \; .< X_N \qquad (3\text{-}4)$$

For a normally distributed population, the arithmetic mean, median, mode and midrange have the same value.

Measures of Dispersion

The second of the two most fundamental measures in statistical analysis is dispersion. Dispersion is a measure of the extent to which the pertinent observations comprising the population are scattered around a measure of central tendency. It may be viewed as a measure of precision or the consistency of, or the variation in, a set of measurements.

The RANGE is the simplest measure of general variability. This is the difference between the highest and lowest value of an entire set of measurements.

$$\text{RANGE} = w = X_N - X_1 \quad \text{Where } X_1 < X_2 <. \; . \; .< X_N \qquad (3\text{-}5)$$

The AVERAGE DEVIATION is the arithmetic mean of the absolute deviation of each value of a set of data from the central value.

$$\text{AVERAGE DEVIATION} = \text{A.D.} = \frac{\sum_{i}^{N} \left| X_i - \bar{X} \right|}{N} \qquad (3\text{-}6)$$

The VARIANCE, or MEAN-SQUARE DEVIATION, is the average of the squared deviations from the mean.

$$\text{VARIANCE} = \sigma^2 = \frac{\sum_{i}^{N} (X_i - \mu)^2}{N} \qquad (3\text{-}7)$$

From a mathematical standpoint the variance is the basic measure of the distribution, but a very frequently used measure of dispersion is the STANDARD DEVIATION, or ROOT MEAN-SQUARE DEVIATION which is the difference between the mean and the point of inflection of a normal curve. The standard deviation is defined as the positive square root of the variance.

$$\text{STANDARD DEVIATION} = \sigma = \sqrt{\sigma^2} \qquad (3\text{-}8)$$

ESTIMATING POPULATION PARAMETERS

A statistical estimation problem involves selecting, on the basis of sample information, an estimate which approximates the value of a population parameter. Estimators are used when practical considerations militate against direct measurement of the population parameter. If the cost of testing exceeds the value of the added benefits, it is uneconomical to measure the parameter directly. If the population is infinite, measurement of all samples is physically impossible. If the test required to measure a particular property alters, consumes or otherwise destroys the product, measurement of all samples is not useful. These considerations apply to testing of bulk petroleum products.

The problem of determining the "best" estimator is varied by the circumstances of the situation. In general, the "best" estimator is one which has a distribution concentrated near the true value of the parameter and which can be applied economically.

Among the statistical criteria for evaluating estimators are unbiasedness, consistency, and efficiency.

The bias of an estimator is the difference between the mean of the distribution of the estimator and the true value of the parameter being estimated. An unbiased estimator then is one which has a distribution having a mean value exactly equal to that of the parameter being estimated.

An estimator is consistent if the probability that an estimate will vary from the true value of the parameter by more than any given amount can be made arbitrarily small by increasing the number of observations in the sample. More simply stated, an estimator is said to be consistent if the reliability of the estimate becomes greater as the sample size is increased.

The efficiency of an estimator is a relative criterion which will be discussed in a later section.

Estimators of the Population Mean

The sample mean, or arithmetic average, is an unbiased estimator of the population mean for any type of population. For a normally distributed population, the sample median and

the sample midrange are also unbiased estimators of the population mean. The purpose of the estimator is to approximate the value of a population parameter, however, the presence of extreme values in a set of sample observations (particularly a small set) could greatly distort the estimate. To minimize distortion, various modifications of the mean, median, and midrange may be computed. These modifications are variously identified in the literature but the majority follow two general patterns;

a. Outlying data in a set are excluded from computation of the mean, median or midrange.

b. An equal number of values from the lower and upper ends of an ordered set are excluded from computation of the mean, or midrange.

The elimination of equal numbers of values from both the high and low ends of the ordered set will not of course change the median. It should also be obvious that the median is a special case of both the symmetrically modified mean and the symmetrically modified midrange. Given a set of six values, the following symmetrically modified means may be generated:

$$\text{Exclude } X_1 \text{ and } X_6 = {}_2\bar{X}_5 = \frac{(X_2 + X_3 + X_4 + X_5)}{4} \qquad (3\text{-}9)$$

$$\text{Exclude } X_1, X_2, X_5, X_6 = {}_3\bar{X}_4 = \frac{(X_3 + X_4)}{2} = \text{Median} \qquad (3\text{-}10)$$

Again using a set of six values, the following symmetrically modified midranges may be generated:

$$\text{Exclude } X_1 \text{ and } X_6 = {}_2C_5 = \frac{(X_2 + X_5)}{2} \tag{3-11}$$

$$\text{Exclude } X_1, X_2, X_5, X_6 = {}_3C_4 = \frac{(X_3 + X_4)}{2} = \text{Median} \tag{3-12}$$

General equation for computation of symmetrically modified mean:

$$(A + 1)\bar{X}(N - A) = \frac{\sum_{A+1}^{N-A} X_i}{(N-2A)} \tag{3-13}$$

Where: A = number of values to be eliminated from each end of the ordered set.

General equation for computation of symmetrically modified midrange:

$$(A + 1)C(N - a) = \frac{X_{(A + 1)} + X_{(N - A)}}{2} \tag{3-14}$$

The principal advantage of arbitrarily discarding data from both ends of an ordered set is the simplicity of the procedure. It has the disadvantage of automatically reducing the effective size of the sample, discarding good data along with any "bad" data. For the most scientifically accurate work, statisticians prefer to discard members of a sample set on an individual basis.[5] This may be limited to

eliminating only those values known to have been influenced by some cause foreign to the rest of the set. It may also be accomplished by following some statistical rule by which values can be discarded with a predetermined error risk. The method of Dixon[6] for testing extreme values, being a nonparametric test, requires only the available sample observations. Dixon's method makes use of critical values of ratios of differences to be expected at various probability levels and for different sample sizes. If the observations in the sample are ranked in order of magnitude as follows:

$$x_1 < x_2 < \ldots < x_{n-1} < x_n$$

the ratio for testing the smallest extreme is:

$$r_{ij} = \frac{x_{1+i} - x_1}{x_{n-j} - x_1} \qquad (3\text{-}15)$$

and the ratio for testing the largest extreme is:

$$r_{ij} = \frac{x_n - x_{n-i}}{x_n - x_{1+j}} \qquad (3\text{-}16)$$

The appropriate ratio for various sample sizes is:

sample size 3 to 7 : r_{10}

sample size 8 to 10 : r_{11}

sample size 11 to 13 : r_{21}

sample size 14 to 30 : r_{22}

Tables giving the maximum expected values for Dixon's ratios are widely reproduced in statistical texts. If an observed ratio exceeds the maximum expected ratio, the extreme value may be rejected with the risk of error set by the tabulated probability level. Another method based on statistical probability is the trial and error method. This method requires an independent estimate of standard deviation. A trial mean is computed from all the observations in the sample. Confidence limits at some reasonable level, say ninety five per cent, are then set around the trial mean. Any extreme data point outside the ninety five per cent confidence interval is assumed not to have come from the same population as the rest of the data and is rejected. A new trial mean and confidence interval are determined based on the remaining data. The entire original set of observations is tested against the new confidence limit and additional data points are rejected and/or previously rejected data points are picked up. The process is repeated until a stable set of values is established, that is, no additional data points are picked up or rejected by the newly computed confidence interval.

Estimators of Population Dispersion

Since the sample mean may not be identical with the population mean, the sum of squares of deviation of the individual sample values from the sample mean will be less

than the sum of squares of deviation of the individual sample values from the population mean. The variance of the sample, computed from the sum of squares of deviation divided by n, the number of items in the sample, will therefore be smaller than if the sum of squares has been calculated from the true population mean. To overcome this bias, the population variance is estimated from a sample by dividing the sum of squares of deviation by n - 1 instead of n.

$$\text{ESTIMATED POPULATION VARIANCE } \hat{\sigma}^2 = s^2 \qquad (3\text{-}17)$$

$$s^2 = \frac{n}{n-1}\,(s^2) = \left[\frac{n}{n-1}\right]\left[\frac{\sum_{i}^{n}(X_i - \bar{X})^2}{n}\right] \qquad (3\text{-}18)$$

An unbiased estimate of the population standard deviation can be obtained by multiplying the square root of the estimated population variance by a correction factor which varies with the type of distribution and the sample size.[7]

$$\text{ESTIMATED POPULATION STANDARD DEVIATION} = \hat{\sigma} = k_n\sqrt{\hat{\sigma}^2} \qquad (3\text{-}19)$$

For a normally distributed population:[8]

$n = 2;\ k_n = 1.253$

$n = 3;\ k_n = 1.128$

$n = 4;\ k_n = 1.085$

$n > 4;\ k_n = 1 \div \frac{1}{4(n-1)}$

The sample range, w, multiplied by the appropriate correction factor forms an unbiased estimator of the population standard deviation. Tables giving correction factors to be applied to the range can be found in readily available textbooks[9] and handbooks[10] and appear to be based on work done by Pearson.[11]

The sample average deviation, A.D., multiplied by a correction factor forms another unbiased estimator of the population standard deviation. Still another, and one which is easier to compute than the average deviation, is the modified linear estimator. Tables of average deviation estimators and modified linear estimators were developed by Dixon and have been published in at least one book which he has co-authored.[12]

Table I summarizes the range, average deviation and modified linear estimators of the population standard deviation for sample sizes two through ten.

TABLE I

UNBIASED ESTIMATORS OF THE POPULATION STANDARD DEVIATION

Sample Size	Range	A.D. from Median	Modified Linear
2	0.8865R	0.8862 DIFA	0.8862 DIFC
3	0.5907R	0.5908 DIFB	0.5908 DIFC
4	0.4857R	0.3770 DIFA	0.4857 DIFC
5	0.4299R	0.3016 DIFB	0.4299 DIFC
6	0.3946R	0.2369 DIFA	0.2619 DIFD
7	0.3698R	0.2031 DIFB	0.2370 DIFD
8	0.3512R	0.1723 DIFA	0.2197 DIFD
9	0.3367R	0.1532 DIFB	0.2068 DIFD
10	0.3249R	0.1353 DIFA	0.1968 DIFD

DIFA = (H-L) where $L = \Sigma X_i$, i = 1 to n/2
and $H = \Sigma X_i$, i = (n/2) + 1 to n

DIFB = (H-L) where $L = \Sigma X_i$, i = 1 to (n-1)/2
$H = \Sigma X_i$, i = (n+3)/2 to n

DIFC = (H-L) where $L = X_1$
$H = X_n$

DIFD = (H-L) where $L = X_1 + X_2$
$H = X_n + X_{(n-1)}$

EFFICIENCY OF ESTIMATORS

The efficiency of an estimator is a relative criterion based on variance. The variance of an estimator is the mean squared deviation of the estimates from the true value of the parameter and the most efficient estimator of a given parameter is the one having the smallest variance. Efficiency is defined as the ratio of the variances of the sampling distributions of the most efficient estimate and the estimate being compared.

$$\text{EFFICIENCY} = E = \frac{\text{Variance of the most efficient estimator}}{\text{Variance of the estimator being compared}}$$

Hence, the efficiency of the most efficient estimator is 1; less efficient estimators have an efficiency of less than 1.

Relative efficiencies are approximately the ratio of sample sizes which will give equal precision in the estimate.[12]

Efficiency of Population Mean Estimators

The sample mean is the efficient estimator of the population mean. The variance of the sampling distribution of the mean is σ^2/n. From the definition of efficiency it follows[13] that the variance of the sampling distribution of an unbiased estimator of the mean of a normal population is σ^2/nE.

The efficiencies of the median and midrange for various sample sizes are given in reference 14. The

efficiency of the median is high for small sample sizes decreasing to a value of 0.637 as n approaches infinity. For the midrange, the efficiency is also high for very small samples but decreases rapidly as the sample size increases, approaching zero as n approaches infinity.

By comparison of the sampling distribution of the means of all possible combinations of two values from a large sample, it can be shown mathematically that the estimator with the highest efficiency among the group is the arithmetic average of the 28.6 percentile value and the 71.4 percentile value.[15] The 25.0 percentile and the 75.0 percentile are usually used in practice for large samples because they are easier to remember and have only a slightly lower efficiency. The limiting efficiency of this modified midrange combination is 0.808 as n approaches infinity. For smaller samples, the efficiency of the Average of the Best Two increases above 0.808. For sample sizes larger than four, the efficiency of the Average of the Best Two as an estimator of the population mean is always greater than that of the median or unmodified midrange. The estimators and efficiencies of the Average of the Best Two for various sample sizes are given in reference 14. Table II gives the estimators based on the Average of the Best Two for samples of size two through ten. It also compares the efficiencies of the median, midrange and Average of the Best Two as estimators of the population mean for these same sample sizes.

TABLE II

EFFICIENCIES OF ESTIMATORS OF THE POPULATION MEAN COMPARED TO THE SAMPLE MEAN

Sample Size	Sample Median	Sample Midrange	Aver. of Best Two Eff.	Aver. of Best Two Estimator
2	1.000	1.000	1.000	$\frac{1}{2}(X_1 + X_2)$
3	0.743	0.920	0.920	$\frac{1}{2}(X_1 + X_3)$
4	0.838	0.838	0.838	$\frac{1}{2}(X_2 + X_3)$
5	0.697	0.767	0.867	$\frac{1}{2}(X_2 + X_4)$
6	0.776	0.706	0.865	$\frac{1}{2}(X_2 + X_5)$
7	0.679	0.654	0.849	$\frac{1}{2}(X_2 + X_6)$
8	0.743	0.610	0.837	$\frac{1}{2}(X_3 + X_6)$
9	0.669	0.572	0.843	$\frac{1}{2}(X_3 + X_7)$
10	0.723	0.539	0.840	$\frac{1}{2}(X_3 + X_8)$

Efficiency of Population Dispersion Estimators

The efficiencies of the range, average deviation and modified linear estimators relative to the square root of s^2 have been determined and published.[16] The efficiency of the range estimator of population standard deviation is relatively high for sample sizes of five or less, but decreases to 0.85 for a sample of size ten and to 0.70 for a sample of size twenty. As the sample size increases indefinitely, it approaches zero. The efficiency of an estimate

based on the average deviation is greater than that of an estimate based on the range for sample sizes larger than six. For sample size ten, it is 0.89. An estimate obtained from the modified linear deviation has an efficiency equal to or greater than either the estimate obtained from the range or the estimate obtained from the average deviation up to sample size five. For larger sample sizes, its efficiency is consistently greater. The efficiencies of the range, average deviation and modified linear estimators for sample sizes two through ten are given in Table III.

TABLE III

EFFICIENCIES OF ESTIMATORS OF POPULATION STANDARD DEVIATIONS AS COMPARED TO S

Sample Size	Range	A.D.	Modified Linear Estimate
2	1.00	1.00	1.00
3	0.99	0.99	0.99
4	0.98	0.91	0.98
5	0.95	0.94	0.96
6	0.93	0.90	0.96
7	0.91	0.92	0.97
8	0.89	0.90	0.97
9	0.87	0.91	0.97
10	0.85	0.89	0.96

CHOOSING STATISTICS AND ESTIMATORS

The proper choice of which statistic or which estimator to use depends upon the problem. Again, the objective is the closest economically obtainable answer to the true value being sought.

Quality surveillance at the command level initially seeks to detect conditions which may require corrective action. Answers which are to be used for management by exception can sacrifice some statistical efficiency for computational efficiency.

Central Tendency

The arithmetic mean is the most widely used measure of central tendency. Perhaps the most important reason for this is that means of samples of uniform size tend to have a normal distribution regardless of the type of distribution of the population from which the samples were drawn. This characteristic of the sample means permits the use of the normal distribution in making probability statements about the population mean with full confidence even if the distribution of the population is unknown or uncertain. The arithmetic mean, being based on all the data, draws the maximum amount of information from the sample. At the same time, it is affected by extreme data, a significant disadvantage when sample size is small and the sample mean is to

be used as an estimate of the population mean. Such is the case when the central tendency value of a correlation test sample distributed among a small number of laboratories is to be used as an estimate of the true value of the property measured. It is obviously important to exclude extraneous values from the computation of the sample mean in such circumstances.

The sample median is a less efficient estimator of the population mean when both the median and the arithmetic mean are computed from the same number of observations. For sample size ten, for example, efficiency of the median is 0.723. The median, however, has the advantage that it is not seriously affected by the retention of extreme values in a sample.[17] Its efficiency in utilizing available data, therefore, is one hundred per cent since none of the observations need be discarded. If, as the result of a test for outliers, three extraneous values were discarded from a set of ten to compute the arithmetic mean estimator of the population mean, the efficiency of utilization of available data is only seventy per cent. An approximation of the relative efficiency of the arithmetic mean and the median as estimators in this case can then be made.

Overall efficiency of arithmetic mean: 0.70 (1.000) = 0.700

Overall efficiency of median: 1.00 (0.723) = 0.723

From this it can readily be seen that the choice of the arithmetic mean as estimator does not guarantee the most efficient estimate in every case.

In the same vein, it must be remembered that although a more efficient estimator has a greater statistical chance of being close to the true population parameter, this does not guarantee that for each sample a more efficient estimate will be closer to the parameter than a less efficient estimate. There is also the question of the relative effort or difficulty in finding the mean value or the median value. If the data are arranged in an order set the median can be located quickly regardless of the sample size. For small samples, say ten or less, the median value can usually be determined by inspection relatively quickly even if the data are not ordered. Mathematically however, the median is hard to handle.

The midrange is a good measure of central tendency for five or less observations but not as good as the mean. For sample sizes larger than five, it is the least efficient estimator of the population mean. Its chief merit is simplicity of calculation but, being the average of the largest and smallest values in a set, it is even more affected by extreme values than the arithmetic mean and the same tests for extreme values are required. However, the midrange is superior to the mean or median for extremely short-tailed

distributions.[18] The Average of the Best Two is a means of artificially creating a short-tailed distribution by chopping off the most widely dispersed values. This estimator offers several advantages. Its construction is such that the probability of being significantly affected by outliers is relatively small, and its efficiency relatively high (0.840 for sample size 10). Yet, it is relatively easy to compute.

Dispersion

The range is the simplest measure of general variability and is very easy to compute. If the sample size is small, say ten or fewer, it is a sensitive measure of the general variability of the population.[19,20] Since only two of the data points are involved in the calculation of the range, it in no way expresses the variation of the other values lying between these two extremes. Therefore, the accuracy of the range estimate of dispersion decreases as sample size increases. None the less, the range is an extremely useful statistic for small samples and is often used in quality control and inspection work.

The average deviation is sensitive to the variability of the population regardless of the size of the sample since it is based on all the data. On the one hand, it is an obviously reasonable measure of variability for small samples

because it is simple to interpret and easy to compute. On the other hand, it is hard to handle in mathematical analysis owing to the use of absolute values.[21] There is a tendency to use the average deviation as a measure of general variability when the median is used as a measure of central tendency because it is a minimum when measured from the median. For a normal distribution, the standard deviation is $\sqrt{\pi/2}$ or 1.253 times the average deviation.[19] If the average deviation is known from historical data, the standard deviation of a measurement can be estimated from this relationship.

The variance and the standard deviation are the most efficient of the estimators of population dispersion. They are harder to compute than the range or the average deviation but are much less affected by extreme values than the range and are mathematically less cumbersome than the average deviation.[22]

CHAPTER IV

ANALYSIS BY NUMERICAL METHODS

INTRODUCTION

The purpose of this thesis as stated in Chapter I is to investigate methods of extracting more definitive information concerning the reliability of the participating laboratories' test results from correlation test data.

Some statistical methods of treating available correlation test data sets which will accomplish this purpose are examined in this chapter. These methods are applied to actual data and the results are interpreted.

The basis of single observation testing is presented first and its limitations are pointed out. Next, a method of analyzing paired sets of data is described and it is shown that two sets of observations are the minimum required to estimate the consistency of a laboratory's results using a proven method. It is also shown that further analysis is possible but is dependent upon an adequate degree of precision being exhibited by the two observations.

A method of treating multiple sets of data follows which is shown to produce a measure of the reliability and a measure of the systematic error of a laboratory's test results as well as an improved measure of the relative

accuracy of results. Two laboratory rating methods are described which could be used as supplements to the Summary of Laboratory Performance described in Chapter I. One method provides an index of accuracy and an index of precision for specific tests. The other provides a laboratory ranking index for the family of tests associated with a given product.

The manner of presentation of each of the methods for analyzing the correlation test data is to discuss the theory and then describe the procedure. The procedural description includes illustrative computations using actual correlation test data obtained from a major military command.

Terminology used in connection with the reliability of laboratory test results is defined in Chapter II.

The statistical measures applied are those discussed in Chapter III. Analysis of laboratory test results is not only a problem of statistical estimation but also a problem of hypothesis testing. The statistical tests applied in this chapter have not themselves been discussed previously in this thesis except for tests of extreme values, but they use the same statistics discussed in Chapter III. The tests will be described as they are introduced into the problem.

TESTING SINGLE OBSERVATIONS

Discussion

Accuracy limits. A minimum of two sets of observations are required to establish an estimate of the precision of a test method. These can be repeated tests by the same operator using the same equipment to establish the operator-equipment precision (repeatability) of the test method, or paired duplicates from separate laboratories to establish the interlaboratory precision (reproducibility) of the method. Once established, the repeatability amount and reproducibility amount can be used to check the accuracy of a single observation when the true value of the property being measured is known or can be estimated.

Let $\bar{d}$ represent the mean difference between pairs of test measurements.

$$\bar{d} = \frac{\sum_{i}^{n} (X_{Aj} - X_{Bj})}{n} \qquad (4\text{-}1)$$

It can be shown that the mean difference between pairs, $\bar{d}$, is $(2/\sqrt{\pi})$ times the standard deviation.[23] By transposing terms, an expression is obtained for computing the standard deviation of a single measurement.

$$\sigma = \frac{\bar{d}\sqrt{\pi}}{2} = 0.8862\bar{d} \qquad (4\text{-}2)$$

A confidence interval to the true value of the property being measured can then be established around the single observation, X_{ij}.

$$\text{Confidence range for } \mu = X \pm z\sigma \qquad (4\text{-}3)$$

Assuming that the single measurement, X_{ij}, comes from a normally distributed population of similar measurements affected by a large number of small random factors, z is the normal deviate appropriate to the desired confidence level. The term, $\pm z\sigma$, is the tolerance set on the precision of measurement X. Therefore, if $\bar{d}$ is known or can be determined, the accuracy of a single measurement can be estimated corresponding to a predetermined degree of confidence.

$$\text{Accuracy limits for } X = \mu \pm z\hat{\sigma}$$

$$= \mu \pm z(0.8862\bar{d}) \qquad (4\text{-}4)$$

The value of z at the five per cent probability level is 1.96.

$$\text{Accuracy limits for } X_{0.95} = \mu \pm (1.96)(0.8862)\bar{d}$$

$$= \mu \pm 1.74\bar{d}$$

These limits can also be expressed as a ninety five per cent accuracy confidence interval for a single observation, X_{ij}.

$$(\mu - z\sigma) \leq X \leq (\mu + z\sigma) \qquad (4\text{-}5)$$

$$(\mu - 1.74\bar{d}) \leq X_{0.95} \leq (\mu + 1.74\bar{d})$$

This interval can then be used to test the hypothesis that the single observation X_{ij} is statistically the same as the true value, μ, of the property being measured.

Procedure

Data and assumptions. The raw data required are the test results for a given property obtained from a single sample which has been divided and distributed among the participating laboratories. Analysis of the data is based upon the following assumptions: (A) The sub-divided samples are homogeneous, that is, there is no quality variation of the material distributed to the various participating laboratories, (B) The universe of observations for each laboratory and all laboratories is normally distributed, (C) The test procedure has been proven, that is, it is adequately described to preclude general misinterpretation of the exact procedures to be followed.

For example, the following single measurements were submitted as the API Gravity of aviation gasoline sample 63-1700 by the ten participating laboratories in a correlation test.

TABLE IV

MEASUREMENTS OF API GRAVITY OF AVIATION GASOLINE
SAMPLE 63-1700 BY TEN LABORATORIES

Test \ Lab.	1	2	3	4	5	6	7	8	9	10
API Grav.	69.8	69.1	69.6	69.1	69.1	69.2	69.2	69.2	69.4	69.2

Decision rule: accuracy. Compute the estimated true API gravity of the gasoline using the sample arithmetic mean as the estimator. Substituting in (3-3):

$$\hat{\mu} = \frac{692.9}{10} = 69.3$$

The ASTM reproducibility amount, R.A., described in Chapter II, can be substituted for the ninety five per cent confidence interval range, $\pm z\sigma$, in (4-4) as a standard to test the statistical accuracy of the single test result obtained by each laboratory. (4-5) then becomes:

$$\left(\hat{\mu} - \frac{R.A.}{2}\right) \leq X_{0.95} \leq \left(\hat{\mu} + \frac{R.A.}{2}\right) \qquad (4\text{-}6)$$

and the decision rule is:

If the observed value is between the estimated population mean minus one half of the ASTM Reproducibility amount and the estimated population mean plus one half the ASTM Reproducibility amount, conclude that results obtained by the laboratory for this test are statistically accurate. If the observed value

lies outside these limits, conclude that results obtained by the laboratory for this test have errors attributable to assignable causes with a five per cent risk of being wrong.

Determine the ASTM Reproducibility amount, R.A., from the Standard Method of Test for API Gravity of Petroleum Products, ASTM Designation: D287-55.[24]

$$R.A. = 0.5 \text{ degrees API}$$

Compute the ninety five per cent confidence limits:

$$\hat{\mu} \pm \frac{R.A.}{2} = 69.3 \pm 0.25$$

At the ninety five per cent confidence level, test the hypothesis that the API Gravity measurement X_j, reported by laboratory j, is statistically the same as the true API Gravity of the sample. Substituting in (4-6):

$$69.05 < X_j < 69.55$$

If the X_j is between 69.05 and 69.55 accept the hypothesis and conclude that results obtained for this test by laboratory j are statistically accurate. If the X_j is less than 69.05 or more than 69.55 reject the hypothesis and conclude that results obtained for this test by laboratory j have errors attributable to assignable causes.

The hypothesis is rejected for two values:

$$X_1 = 69.8 > 69.55$$

$$X_3 = 69.6 > 69.55$$

The distorting effect of outlying data on estimates of population parameters was discussed in Chapter III and a trial and error method of eliminating outliers from computation of the mean was described. Applying this method:

$$\hat{\mu} = \frac{\sum_{j}^{10} X_j - X_1 - X_3}{8} = \frac{553.5}{8} = 69.2$$

Compute new ninety five per cent confidence limits:

$$\hat{\mu} \pm \frac{R.A.}{2} = 69.2 \pm 0.25$$

Substitute in equation (4-6) and retest the hypothesis for all ten measurements X_j:

$$68.95 < X_j < 69.45$$

The hypothesis is rejected for the same two values:

$$X_1 = 69.8 > 69.45$$

$$X_3 = 69.6 > 69.45$$

Since no additional data points were rejected and none previously rejected were picked up, a stable set of values has been determined.

This is the method presently used to evaluate correlation test results. It has been previously pointed out that this method gives no indication of whether systematic errors or mistakes are the causes of out-of-control observations.

TESTING PAIRED OBSERVATIONS

Discussion

Just as a minimum of two sets of observations were required to establish an estimate of the precision of a test method, two observations are the minimum data required to estimate the consistency of a laboratory's results using a proven method.

Precision Limits

Two observations can be analyzed for precision by estimating the standard deviation from the mean difference between pairs. Precision limits for μ:

$$\mu = X \pm z\sigma \tag{4-3}$$

Confidence interval for X:

$$(\mu - z\sigma) \leq X \leq (\mu + z\sigma) \tag{4-5}$$

Let the confidence range, $\pm z\sigma$, which is constant for a given probability level, be represented by the symbol 2C. The paired test results from one laboratory are represented by X_{Aj} and X_{Bj}. The sample mean, $\overline{X}_j$, is the estimator of the population mean. Then:

$$(\overline{X}_j - C) \leq X_{Aj} \leq (\overline{X}_j + C) \tag{4-7}$$

Substituting for $\overline{X}_j$:

$$\left[\frac{X_{Aj} + X_{Bj}}{2}\right] - C \leq X_{Aj} \leq \left[\frac{X_{Aj} + X_{Bj}}{2}\right] + C \tag{4-8}$$

Clearing fractions:

$$(X_{Aj} + X_{Bj} - 2C) \leq (2X_{Aj}) \leq (X_{Aj} + X_{Bj} + 2C)$$

Subtracting X_{Aj}:

$$(X_{Bj} - 2C) \leq (X_{Aj}) \leq (X_{Bj} + 2C)$$

Subtracting X_{Bj}:

$$(-2C) \leq (X_{Aj} - X_{Bj}) \leq (+2C)$$

Transposing:

$$(X_{Aj} - X_{Bj}) \leq \pm 2C \tag{4-9}$$

Therefore:

$$\left| X_{Aj} - X_{Bj} \right| \leq 2C \leq 2z\sigma \tag{4-10}$$

Likewise:

$$(\bar{X}_j - C) \leq X_{Bj} \leq (\bar{X}_j + C)$$

Substituting for $\bar{X}_j$, clearing fractions, subtracting X_{Aj} and X_{Bj}, and transposing terms:

$$(X_{Bj} - X_{Aj}) \leq \pm 2C$$

And:

$$\left| X_{Bj} - X_{Aj} \right| \leq 2C$$

But:

$$\left| X_{Bj} - X_{Aj} \right| = \left| X_{Aj} - X_{Bj} \right|$$

Therefore:

$$\left| X_{Aj} - X_{Bj} \right| \leq 2C \leq 2z\sigma \tag{4-11}$$

at the ninety five per cent confidence level:

$$\left|X_{Aj} - X_{Bj}\right| \leq 2(1.74)\bar{d} \leq 3.48\bar{d}$$

This range limit can then be used to test the hypothesis that a pair of observations (X_{Aj}, X_{Bj}) are statistically one and the same value. If they are, further statistical inferences may be drawn from them.

Estimating systematic error. If the two observations from a laboratory show an acceptable degree of precision, an estimate can be made of the amount and direction of systematic error or bias which they contain.

$$\text{BIAS} = \frac{(X_{Aj} - \bar{X}_A) + (X_{Bj} - \bar{X}_B)}{2} \qquad (4\text{-}12)$$

or, for simpler calculation,

$$\text{BIAS} = \frac{(X_{Aj} + X_{Bj})}{2} - \frac{(\bar{X}_A + \bar{X}_B)}{2} \qquad (4\text{-}13)$$

$$= \bar{X}_j - \bar{\bar{X}}$$

Although constant factors may be present in measurements which are not statistically precise, there is a high probability that either or both of the measurements also contain errors caused by mistakes of unknown magnitude and direction. A 'bias' computation would be meaningless in such circumstances, could only cause confusion and should not be made.

Accuracy limits. The standard deviation of the means of samples of size n is estimated by dividing the estimated population standard deviation by the square root of n. One possible way to define accuracy in a normally distributed population is:

$$\bar{X} \leq \mu \pm z\left(\frac{\sigma}{\sqrt{n}}\right) \qquad (4\text{-}14)$$

where:

μ = the true value of the property being measured

σ = the population standard deviation

n = the sample size

z = the normal deviate for the desired level of confidence

$\bar{X} = \frac{\Sigma X}{n}$

But, once again, the proper choice of a statistic or estimator is dependent upon the available data and the intended purpose for which it is to be used. For small samples acceptance of the hypothesis that the sample mean and the true value of the property being measured are statistically one and the same value on the basis of the above test may occur even when the situation is not true. For example, assume two observations are obtained of a property whose true value is zero. It can be readily seen that regardless of magnitude, if the two observations have the same value but opposite signs the average will be zero. Although the mean of the

observations would be statistically "accurate" the hypothesis test is obviously meaningless for such a situation.

One simple way to handle this dilemma is to tie the accuracy determination to the precision test. The hypothesis test for accuracy of a laboratory's test results would then be modified to the extent that $\bar{X}$ would be redefined as the average of a set of laboratory test results which are statistically precise at some specified level.

Procedure

Data. The raw data required are the test results for a given property obtained from two samples which have been divided and distributed among the m participating laboratories. It is not necessary that both samples be of the same product. It may be feasible to pool test results of different products. Volk states that, in comparing paired data, the pairs do not have to be measures of the same thing, but the individual measurements in a pair will be made at the same conditions.[25] The objective is to avoid introducing additional sources of variability. Generally, this objective can be accomplished if the test procedures are identical and if the samples are reasonably close in the magnitude of the property being evaluated.[26] However, even though pooled test results are obtained from statistically homogeneous samples, if they are not duplicate tests of the

same sample, they do not have a common mean. Consequently, the observations X_{1j} and X_{2j} cannot be compared directly. The algebraic deviation from the mean, v_{ij}, for each observation must be determined by subtracting the mean, $\bar{X}_i$, computed for each test from each observation reported for that test.

$$v_{ij} = X_{ij} - \bar{X}_i \qquad (4\text{-}15)$$

More will be said about the pooling of data to form larger samples in the section on multiple test results.

Correlation test results of the ten per cent distillation point of two different samples of aviation gasoline, grade 115/145 provided the data which will be used to illustrate the procedure. These two sets of values are given in Table V. The results labeled as Test 1 are measurements taken on correlation test sample 64-27. Those labeled as Test 2 are measurements taken on correlation test sample 64-3599. The corresponding matrix of observations, v_{ij}, is given in Table VI.

TABLE V

DISTILLATION OF AVIATION GASOLINE GRADE 115/145 10 PER CENT RECEIVED @ °F

Laboratory j	Test i 1	2
1	152	148
2	148	146
3	149	147
4	147	144
5	150	142
6	150	148
7	148	146
8	145	148
9	147	149
10	146	150
$\bar{X}_i$	148.2	146.8

Test 1: Sample 64-27

Test 2: Sample 64-3599

TABLE VI

ANALYSIS OF TEST RESULTS: DISTILLATION OF AVIATION GASOLINE GRADE 115/145. DEVIATION FROM THE MEAN: 10 PER CENT RECEIVED @ °F

Lab. j	Test i 1	2	$\lvert v_{1j}-v_{2j}\rvert$	Bias $\bar{v}_j$
1	+3.8	+1.2	2.6	+2.5
2	-0.2	-0.8	0.6	-0.5
3	+0.8	+0.2	0.6	+0.5
4	-1.2	-2.8	1.6	-2.0
5	+1.8	-4.8	6.6	-1.5
6	+1.8	+1.2	0.6	+1.5
7	-0.2	-0.8	0.6	-0.5
8	-3.2	+1.2	4.4	-1.0
9	-1.2	+2.2	3.4	+0.5
10	-2.2	+3.2	5.4	+0.5

Test 1: Sample 64-27

Test 2: Sample 64-3599

Assumptions. Analysis of the data is based upon the following assumptions: (A) The sub-divided samples are homogeneous, that is, there is no quality variation of the material distributed to the various participating laboratories for each test, (B) The universe of observations for each activity and all activities is normally distributed; and, (C) The test procedure has been proven, that is, it is adequately described to preclude general misinterpretation of the exact procedures to be followed.

Decision rule: precision. The ASTM reproducibility amount, R.A., described in Chapter II, can be substituted for the ninety five per cent confidence interval range 2C in (4-11) as a standard to test the statistical precision of the pair of test results obtained by each laboratory. (4-11) then becomes:

$$\left| v_{1j} - v_{2j} \right| \leq R.A. \qquad (4\text{-}16)$$

and the decision rule is:

> If the absolute value of the difference between the deviation from the test means of two independent measurements is equal to or less than the ASTM reproducibility amount for the test, conclude that results obtained by the laboratory for this test are sufficiently precise, i.e., errors affecting results are probably due to chance causes inherent to the prescribed test method. If the absolute difference is

greater than the ASTM reproducibility amount, conclude that results obtained from performance of this test by the laboratory have errors attributable to assignable causes with a five per cent risk of being wrong.

Determine the ASTM reproducibility amount, R.A., from the Standard Method of Test for Distillation of Petroleum Products, ASTM Designation: D86-61.[27]

$$R.A. = 7\ ^{o}F$$

For the m laboratories, compute:

$$\left| v_{1j} - v_{2j} \right| \quad , \quad j = 1 \text{ to } m$$

At the ninety five per cent confidence level, test the hypothesis that the ten per cent distillation point measurements X_{1j} and X_{2j} reported by laboratory j are statistically the same in respect to their deviation from the true values of the ten per cent distillation points of samples 1 and 2 respectively. Substituting in (4-16):

$$\left| v_{1j} - v_{2j} \right| \leq 7$$

If the absolute difference between v_{1j} and v_{2j} is equal to or less than 7, accept the hypothesis and conclude that results obtained for this test by laboratory j are sufficiently precise. If the difference is greater than 7, reject the hypothesis and conclude that results obtained for this test by laboratory j fail to meet minimum standards for precision.

The differences, $|v_{1j} - v_{2j}|$, for the illustrative test results are tabulated in Table VI. The paired test results from all laboratories are precise according to the established standard. Consequently, all may be further analyzed for average bias and for accuracy.

Bias measurement. The mean deviation from the mean of the paired test results reported by laboratory j is determined by:

$$\bar{v}_j = \frac{v_{1j} + v_{2j}}{2} \qquad (4\text{-}17)$$

This is equivalent to (4-12) for the bias estimate based on two observations from a laboratory which shows an acceptable degree of precision. The values $\bar{v}_j$ computed from the illustrative data appear in Table VI. These values will be further utilized in testing the accuracy of the laboratories.

Decision rule: accuracy. A test for accuracy is given by (4-14) in which $\bar{X}$ is defined as the average of a set of laboratory test results which are statistically precise at some specified level. Substituting $\bar{v}_j$ for $\bar{X}$ and $\bar{\bar{v}}_{ij}$ for μ, (4-14) becomes:

$$\bar{v}_j \leq \bar{\bar{v}}_{ij} \pm \frac{z\sigma}{\sqrt{n}} \qquad (4\text{-}18)$$

But $\bar{\bar{v}}_{ij}$ is zero by definition. Therefore (4-18) becomes:

$$\bar{v}_j \leq 0 \pm \frac{z\sigma}{\sqrt{n}} \qquad (4\text{-}19)$$

The ASTM reproducibility amount, R.A., can be substituted for the ninety five per cent confidence interval range $\pm z\sigma$ in (4-19) as a standard to test the statistical accuracy of the paired test results obtained by each laboratory. Also substituting for n, (4-19) becomes:

$$-\frac{R.A.}{2\sqrt{2}} \leq \bar{v}_j \leq +\frac{R.A.}{2\sqrt{2}} \qquad (4\text{-}20)$$

and the decision rule is:

> If two single observations obtained from statistically homogeneous sources are statistically precise at the ninety five per cent level, and if the absolute value of the average variation from the mean of the paired single observations is within the ninety five per cent confidence range based on the applicable ASTM Reproducibility amount, conclude that results obtained by the laboratory for this test are accurate. If the absolute value of the average is above or below the ninety five per cent confidence range, conclude that the results obtained performing this test contain errors which cannot be accounted for by chance causes with a five per cent risk of having reached the wrong conclusion.

Compute the ninety five per cent confidence limits:

$$\pm \frac{R.A.}{2\sqrt{2}} = 0 \pm \frac{7}{2\sqrt{2}} = \pm 2.5$$

At the ninety five per cent confidence level, test the hypothesis that in regard to deviation from the true value of the property measured, the average of a pair of measurements is statistically the same as zero. Substituting in (4-20):

$$-2.5 \leq \bar{v}_j \leq +2.5$$

If the $\bar{v}_j$ is between -2.5 and +2.5 accept the hypothesis and conclude that, on the average, results obtained for this test by laboratory j are sufficiently accurate. If the $\bar{v}_j$ is less than -2.5 or greater than +2.5, reject the hypothesis and conclude that, on the average, the results obtained for this test by laboratory j have errors attributable to assignable causes.

The hypothesis is accepted for the ten laboratories in the example but laboratory 1 is on the borderline.

While these results produce a quick and satisfactory indication of accuracy, they do not make full use of the available information. They do not take into consideration the probability of statistically independent events. The outcome of either of two separate laboratory tests is not conditioned by the outcome of the other. Therefore observation A and observation B are statistically independent and the probability of both A and B occurring is the product of the probability of A occurring and the probability of B occurring.

$$Pr(A \text{ and } B) = (Pr\ A)(Pr\ B) \qquad (4\text{-}21)$$

The hypothesis test employed assumes that both observations (either the X_{Aj} and X_{Bj} replicate measurements or the v_{Aj} and v_{Bj} single measurements) come from the same normally distributed population. Therefore the distance from the population mean of each observation can be expressed in terms of multiples of the population standard deviation, that is, the normal deviate, z. The area under the frequency distribution curve, bounded by the interval dz which includes z_A, measures the probability of obtaining observation A in a random sample as shown in Figure 4-1. Likewise, the area under the frequency distribution curve bounded by the interval dz which includes z_B measures the probability of obtaining observation B in a random sample. In a normal distribution

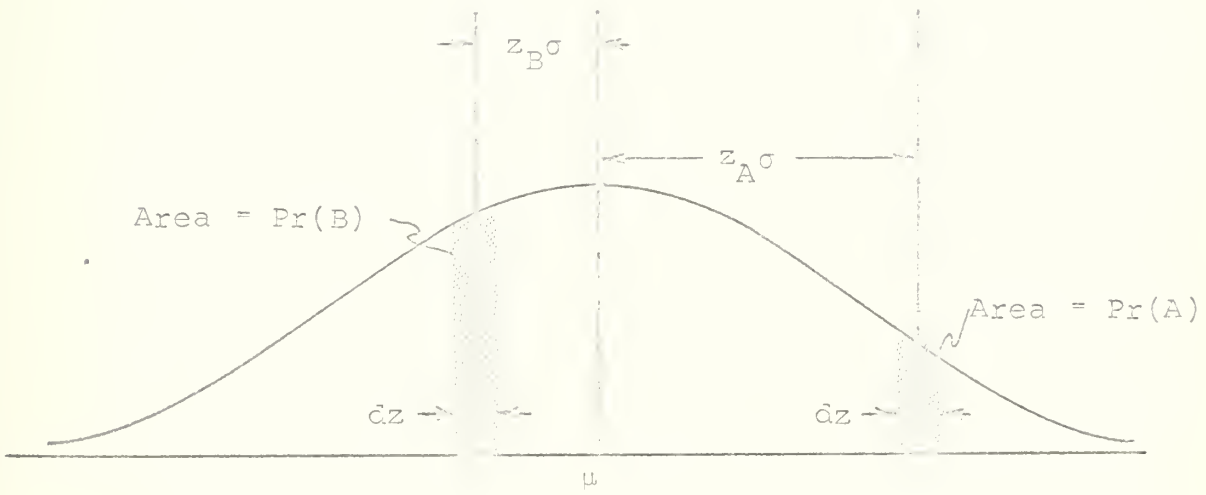

FIGURE 4-1

THE PROBABILITY OF OBTAINING A GIVEN VALUE FROM A NORMAL DISTRIBUTION

the probability of obtaining a particular value of z diminishes as z increases. Therefore, the probability of obtaining two observations out in one or the other tail of the distribution due to chance causes alone is very small. Conversely, the probability of obtaining two observations close to the population mean if only chance causes are affecting the measurements is relatively high.

Given two sets of results, ($z_{A1} = 1.96$, $z_{B1} = 0.0$) and ($z_{A2} = 1.96$, $z_{B2} = 1.96$) one would conclude intuitively that results from laboratory 1 are more apt to be accurate than results from laboratory 2. Indeed it can be shown that if a finite z-interval of 0.02 is substituted for dz, the probability of obtaining the subset of measurements (z_{A1}, z_{B1}) due to random variation is more than six and a half times as great as the probability of obtaining the subset (z_{A2}, z_{B2}).

$A1 = Pr\ (1.95 < z < 1.97) = .0012$

$B1 = Pr\ (-.01 < z < +.01) = .0080$

$Pr\ (A1 \text{ and } B1) = (.0012)(.0080) = 9.60(10^{-6})$

$A2 = Pr\ (1.95 < z < 1.97) = .0012$

$B2 = Pr\ (1.95 < z < 1.97) = .0012$

$Pr\ (A2 \text{ and } B2) = (.0012)(.0012) = 1.44(10^{-6})$

The consequences of applying this rule do not appear to be significant enough to justify the considerable extra effort required. However the overall effect should be noted.

Viewed from the standpoint of confidence level, the probability of an observation A greater than $z_{0.95}$ and an observation B greater than $z_{0.95}$ is (0.05)(0.05) or 0.0025. Therefore the decision rule carries a risk which varies from 0.05 to 0.0025 of wrongly classifying an "accurate" activity as "inaccurate." Conversely, the risk of failing to detect an "inaccurate" activity is increased.

TESTING MULTIPLE OBSERVATIONS

Discussion

Consider the results of n tests submitted by m laboratories as represented by the matrix of Table VII. Assume that the universe of observations for each test is normally distributed. The objective is to determine the kind and magnitude of variability that can be expected to be included in observations made by a given laboratory. Since the measurement quality of interest is variability, the first step is to convert the data to measurements of variation or algebraic distance from the true value of the property being measured.

For each test, a sample mean, $\bar{X}_i$, can be obtained which can be used as an estimator of the population mean. If the n tests were duplicate tests of homogeneous samples taken from the same population, the test means would be expected to cluster around a single value, the population

TABLE VII

SYMBOLIC MATRIX OF RESULTS OF n TESTS
SUBMITTED BY m LABORATORIES

Test i \ Lab. j	1	2	. . .	j	. . .	m
1	X_{11}	X_{12}	. . .	X_{1j}	. . .	X_{1m}
2	X_{21}	X_{22}	. . .	X_{2j}	. . .	X_{2m}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
i	X_{i1}	X_{i2}	. . .	X_{ij}	. . .	X_{im}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
n	X_{n1}	X_{n2}	. . .	X_{nj}	. . .	X_{nm}

mean, μ. The average mean, $\bar{\bar{X}}$, becomes a better estimator of the population mean which can be used to determine the algebraic variation from the mean, v_{ij}, of each of the n times m observations. If the n tests were not duplicate tests of the same batch of product, but (A) the tests were identical in procedure, and (B) the materials tested are close enough in magnitude of the property measured as to preclude any significant variation in the random error due to material, the test results can be compared in regard to variation from

the mean but do not have a common mean. The v_{ij} for each observation can be determined only by subtracting the $\bar{X}_i$ computed for each test from each observation reported for that test. A new matrix, Table VIII, results.

TABLE VIII

SYMBOLIC MATRIX OF DEVIATION, v_{ij}, FROM ESTIMATED TEST POPULATION MEAN

Test \ Lab.	1	2	. . .	j	. . .	m
1	v_{11}	v_{12}	. . .	v_{1j}	. . .	v_{1m}
2	v_{21}	v_{22}	. . .	v_{2j}	. . .	v_{2m}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
i	v_{i1}	v_{i2}	. . .	v_{ij}	. . .	v_{im}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
n	v_{ni}	v_{n2}	. . .	v_{nj}	. . .	v_{nm}

Homogeneity of Variance

By pooling data sets in this manner, larger samples are available for estimating the variability of laboratory observations resulting in potentially better estimates. Only data sets having statistically homogeneous variances are really comparable, however. A statistical test was devised

by Bartlett for passing judgement in such cases. If n sets of data are available with varying numbers of observations, m, in each set, the statistical parameter, B, can be computed in the following manner:

TABLE IX

BARTLETT'S TEST FOR HOMOGENEITY OF VARIANCES

Test Data Set	s_i^2	Degrees of Freedom $f_i=(m_i-1)$	$f_i s_i^2$	$\ln s_i^2$	$f_i \ln s_i^2$	$\frac{1}{f_i}$
1	s_1^2	f_1	$f_1 s_1^2$	$\ln s_1^2$	$f_1 \ln s_1^2$	$1/f_1$
2	s_2^2	f_2	$f_2 s_2^2$	$\ln s_2^2$	$f_2 \ln s_2^2$	$1/f_2$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	s_i^2	f_i	$f_i s_i^2$	$\ln s_i^2$	$f_i \ln s_i^2$	$1/f_i$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	s_n^2	f_n	$f_n s_n^2$	$\ln s_n^2$	$f_n \ln s_n^2$	$1/f_n$
TOTALS		f	$\Sigma f_i s_i^2$		$\Sigma f_i \ln s_i^2$	$\Sigma \frac{1}{f_i}$

Compute:

$$s^2 = \frac{\Sigma f_i s_i^2}{f} \qquad (4\text{-}22)$$

and:

$$f \ln s^2 \qquad (4\text{-}23)$$

then:

$$B = \frac{1}{C} \left(f \ln s^2 - \Sigma f_i \ln s_i^2 \right) \qquad (4\text{-}24)$$

The value of B may be computed initially without evaluating the correction factor, C. The critical value of B at the selected confidence level may be read from a statistical

table of chi square available in most statistics texts and handbooks, entering the table with (n-1) degrees of freedom. If B is significant at the selected confidence level, i.e., exceeds the critical value, it may then be divided by the correction factor, C, computed as follows:

$$C = 1 + \frac{\Sigma \frac{1}{f_i} - \frac{1}{f}}{3(n-1)} \qquad (4\text{-}25)$$

If the corrected value of B is also significant at the selected confidence level, reject the hypothesis that the sets of data being compared have the same variance.

Analyzing the Data

For each of the m participating laboratories, an average algebraic variation from the mean, $\bar{v}_j$, can be computed. This is the average accuracy error and constitutes a point estimate of the magnitude and direction of the systematic error or bias.

An estimated population variance, $\hat{\sigma}_j^2$, also can be compuated for each activity, using S^2_j as the estimator. This is a measure of the variation in the point estimate of the systematic error due to random and accidental causes.

Having sufficiently isolated random, systematic, and accidental errors to obtain an approximate measure of each, a judgement can be made concerning laboratory reliability,

by comparing the measures of reliability for each activity against matching standards.

Procedure

Data and assumption. The raw data required are the test results for a given property obtained from n samples of different batches of product which have each been divided and distributed among the m participating laboratories. It is not necessary that all samples be of the same product. It may be feasible to pool test results of different products. The considerations in this regard are the same as for paired data. When doubt exists, a statistical test for homogenity of variance of the pooled data is appropriate.

Analysis of the data is based upon the same assumptions already stated for paired data.

To illustrate the procedure, the correlation test results used are the measurements of API Gravity for five different products. The matrix of these observations is given in Table X.

Estimating the population mean. Since the several sets of test results are not repeat measurements of the same product sample, the tests do not have a common mean. A separate estimate of the population mean, μ_i, must be made for each of the i tests.

TABLE X

API GRAVITY OF FIVE PRODUCTS MEASURED BY TEN LABORATORIES

Test i \ Lab. j	1	2	3	4	5	6	7	8	9	10	$\bar{X}_i$	Aver. of Best Two
1	69.2	68.2	68.8	68.3	68.2	68.1	68.4	68.1	68.3	68.6	68.5	68.4
2	59.6	59.5	59.9	59.4	59.6	59.6	59.6	59.6	59.6	59.6	59.6	59.6
3	55.8	55.7	55.8	55.5	55.5	55.4	55.5	55.6	55.8	55.8	55.6	55.6
4	38.1	37.9	38.2	38.1	38.1	38.1	38.1	38.1	38.0	38.1	38.1	38.1
5	20.2	20.3	20.0	20.1	20.2	20.1	20.1	20.2	21.2*	20.3	20.2	20.2

i	Product	Sample
1	Aviation Gasoline, Grade 115/145	63-01
2	Combat Automotive Gasoline	63-02
3	Jet Fuel, Grade JP-4	63-04
4	Marine Diesel Fuel Oil, Type I	63-03
5	Lubricating Oil, Heavy Duty, Grade 30	63-05

The most efficient estimator of the population mean is the sample arithmetic mean. Because outliers can have a significant effect on the arithmetic mean of small samples, an appropriate test should be applied to any values which appear extreme. Dixon's test for extreme values, described in Chapter II, will be used to check the two doubtful values in Table X:

$$X_{11} = 69.2 \text{ and } X_{59} = 21.2$$

Ratio test symbol r_{11} for the largest extreme applies in both cases. The critical value for test r_{11} at the 0.05 level is 0.477.

Check $X_{11} = 69.2$:

$$\frac{x_{10} - x_9}{x_{10} - x_2} = \frac{69.2-68.8}{69.2-68.1} = \frac{0.4}{1.1} = 0.364$$

Since the ratio does not exceed the critical value of 0.477 accept the hypothesis that X_{11} comes from the same population as the other results submitted for Test 1.

Check $X_{59} = 21.2$:

$$\frac{21.2 - 20.3}{21.2 - 20.1} = \frac{0.9}{1.1} = 0.811$$

Since the ratio exceeds the critical value of 0.477 reject the hypothesis that X_{59} comes from the same population as the other results submitted for Test 5. An asterisk is used to flag X_{59} as an outlier in the tabulated data.

Compute $\bar{X}_1$ for each of the n tests which have been pooled for the analysis and use these values as estimates of the corresponding population means. When computing the mean for Test 5, exclude X_{59} from the computation to minimize the probability of distorting the estimated true API Gravity of sample 63-05. The arithmetic mean estimate of the true API Gravity for each of the five tests is tabulated in column $\bar{X}_i$ of Table X.

To avoid the necessity of testing for outliers, it may be desired to use the Average of the Best Two rather than the sample arithmetic mean as the estimator of the population mean. This estimator, discussed in Chapter III, is relatively easy to compute and has a high efficiency for small sample sizes. For sample size 10,

$$\bar{X} = \text{Aver. of Best Two} = \tfrac{1}{2}(x_3 + x_8) \qquad (4\text{-}26)$$

Where: x_3 = the X_{ij} ranking third in magnitude among observations for test i.

x_8 = the X_{ij} ranking eighth in magnitude among observations for test i.

The Average of the Best Two estimate of the true API Gravity is also tabulated in Table X for comparison with the arithmetic mean estimate. Both values are identical for four of the tests and are separated by only 0.1 degree API for Test 1.

Computing the matrix of deviations from the mean. Subtract $\bar{X}_i$ from each of the m observations submitted for test i to obtain the values v_{ij} which measure the algebraic deviation of each observation from the estimated population mean.

$$v_{ij} = X_{ij} - \bar{X}_i \tag{4-15}$$

The resulting matrix of values for the illustrative tests is given in Table XI.

Testing for homogeneity of variance. Determine the estimated population variance for each test using an unbiased estimator.

$$\hat{\sigma}^2 = s^2 \tag{3-17}$$

$$s_i^2 = \frac{1}{(m-1)} \left[\sum_j^m (v_{ij} - \bar{v}_i)^2 \right] \tag{3-18}$$

A simpler computational form is:

$$s_i^2 = \frac{1}{m(m-1)} \left[m \sum_j^m v_{ij}^2 - \left(\sum_j^m v_{ij} \right)^2 \right] \tag{4-27}$$

The values, s_i^2, of the estimated population variance for each of the five illustrative tests are given in Table XII.

Again it may be desired to use a short-cut method of computation. The Modified Linear Estimator of the population standard deviation described in Chapter III was characterized as being relatively easy to compute and having a high efficiency. For Tests 1 through 4 the Modified Linear Estimator for sample size 10 is:

TABLE XI

DEVIATION, v_{ij}, FROM TEST MEAN, μ_i, API GRAVITY OF FIVE PRODUCTS MEASURED BY TEN LABORATORIES

Test i \ Lab. j	1	2	3	4	5	6	7	8	9	10
1	+0.7	-0.3	+0.3	-0.2	-0.3	-0.4	-0.1	-0.4	-0.2	+0.1
2	0.0	-0.1	+0.3	-0.2	0.0	0.0	0.0	0.0	0.0	0.0
3	+0.2	+0.1	+0.2	-0.1	-0.1	-0.2	-0.1	0.0	+0.2	+0.2
4	0.0	-0.2	+0.1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0
5	0.0	+0.1	-0.2	-0.1	0.0	-0.1	-0.1	0.0	+1.0	+0.1

TABLE XII

BARTLETT'S TEST FOR HOMOGENEITY OF VARIANCE OF FIVE TESTS OF API GRAVITY

Test i	$\sum_j^m v_{ij}$	$(\sum v_{ij})^2$	$\sum_j^m v_{ij}^2$	S_i^{2*}	$f_i = (m-1)$	$f_i(S_i^2)$	$\ln S_i^2$	$f_i(\ln S_i^2)$	$1/f_i$	S_i^2 by Modified Linear Estimator
1	[illegible]	0.64	1.12	0.117	9	1.053	-2.45	- 21.05	0.111	0.113
2	0.0	0.0	0.14	0.016	9	0.144	-4.14	- 37.26	0.111	0.014
3	+0.4	0.16	0.24	0.025	9	0.225	-3.69	- 33.21	0.111	0.014
4	-0.2	0.04	0.06	0.006	9	0.054	-5.12	- 46.08	0.111	0.006
5	-0.3	0.09	0.09	0.010	8	0.080	-4.60	- 36.80	0.125	0.007
TOTALS					44	1.556		-174.40	0.569	

* $$S_i^2 = \frac{\sum_j^m (v_{ij} - \bar{v}_i)^2}{(m-1)} = \frac{1}{m(m-1)} \left[m \sum_j^m v_{ij}^2 - \left(\sum_j^m v_{ij} \right)^2 \right]$$

$$\hat{\sigma}_i = 0.1968 \, (x_{10} + x_9 - x_1 - x_2) \qquad (4\text{-}28)$$

Where: x_1, x_2, x_9 and x_{10} are the first, second, ninth and tenth values ranked in order of magnitude from smallest to largest. Extreme values have a significant effect on estimates computed from the Modified Linear Estimator. Observation X_{59} should therefore be excluded from the computation of the estimated population standard deviation of Test 5, reducing the sample size to 9. The Modified Linear Estimator for sample size 9 is:

$$\hat{\sigma}_i = 0.2068 \, (x_9 + x_8 - x_1 - x_2) \qquad (4\text{-}29)$$

Squaring the estimate of population standard deviation obtained from these computations gives an estimate of the population variance of each of the five tests. The results are tabulated in Table XII for comparison with the efficient estimator computed by equation (3-18). Agreement is reasonably close except for Test 3. If this estimator is used in connection with Bartlett's test it is recommended that any borderline indications of homogeneity or non-homogeneity of variance be rechecked using the efficient estimator of the population variance.

Compute s^2 from (4-27):

$$s^2 = \frac{1.556}{44} = 0.0354$$

Then:

$$f(\ln s^2) = 44 \, (-3.34) = -147.00$$

Compute B from (4-24) without evaluating the correction factor, C:

$$B = \frac{1}{C}\left[-147.00 - (-174.40)\right] = \frac{1}{C}(27.40)$$

Refer to a statistical table of chi-square. Enter the table with (n-1) = 4 degrees of freedom to determine the critical value at the ninety five per cent confidence level.

$$\chi^2 = 9.488$$

The value of B exceeds the critical value indicating that there is a significant difference among the variances of the five sets of test data.

Compute correction factor, C, from (4-25):

$$C = 1 + \frac{0.569 - 0.023}{3(5-1)} = 1.045$$

Determine the corrected value of B:

$$B = \frac{27.40}{1.045} = 26.20$$

Since B still exceeds the critical value at the ninety five per cent confidence level, reject the hypothesis that the five sets of test results have the same variance and conclude that they cannot be pooled to form a single large sample.

Form a subset of four tests by dropping the set exhibiting the most extreme variance which is Test 1. Test this subset for homogeneity of variance.

$$f = \Sigma f_i = 35$$

$$\Sigma f_i(s_i^2) = 0.503$$

$$\Sigma f_i(\ln s_i^2) = -153.35$$

$$s^2 = \frac{0.503}{35} = 0.0144$$

$$f(\ln s^2) = (35)(\ln 0.0144) = -147.70$$

$$B = \frac{1}{C}[-147.70 - (-153.35)] = \frac{1}{C}\ 5.65$$

Entering a table of chi-square with (n-1) = 3 degrees of freedom, determine the critical value at the ninety five per cent confidence level.

$$\chi^2 = 7.815$$

Since the value of B is less than the critical value, accept the hypothesis that the four sets of test results have the same variance and conclude that they are comparable and can be pooled. The new matrix is given in Table XIII.

Estimating bias. Compute the average algebraic deviation from the mean, $\bar{v}_j$ for each of the j activities, excluding outliers from the computation. The $\bar{v}_j$ can then be used as a point estimate of the magnitude and direction of the bias in results reported for this type of test by laboratory j. The reason for excluding the extreme values is that they were previously rejected on the basis of a hypothesis test leading to decisions that they probably contained errors due to mistakes. Inclusion of these mistakes would distort the bias.

TABLE XIII

ANALYSIS OF API GRAVITY BY TEN LABORATORIES FOR FOUR TESTS HOMOGENEOUS AT THE 95 PER CENT CONFIDENCE LEVEL BY BARTLETT'S TEST

Sample	Test i \ Lab. j	1	2	3	4	5	6	7	8	9	10
63-02	1	0.0	-0.1	+0.3	-0.2	0.0	0.0	0.0	0.0	0.0	0.0
63-04	2	+0.2	+0.1	+0.2	-0.1	-0.1	-0.2	-0.1	0.0	+0.2	+0.2
63-03	3	0.0	-0.2	+0.1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0
63-05	4	0.0	+0.1	-0.2	-0.1	0.0	-0.1	-0.1	0.0	+1.0*	+0.1
Bias Estimate = $\bar{v}_j$ excluding extreme values:		+0.05	-0.02	+0.10	-0.10	-0.02	-0.08	-0.05	0.0	+0.02	+0.08
Accuracy Coefficient, $A.C._j = \lvert \bar{v}_j \rvert$ including all data:		0.05	0.02	0.10	0.10	0.02	0.08	0.05	0.0	0.25	0.08
Accuracy Index, $A.I._j = (0.125/A.C._j) - 1.0$		+1.5	+5.2	+0.2	+0.2	+5.2	+0.6	+1.5	∞	-0.5	+0.6

TABLE XIII (continued)

Sample	Test i \ Lab. j	1	2	3	4	5	6	7	8	9	10
	$\sum_{i}^{n} v_{ij}$	+0.2	-0.1	+0.4	-0.4	-0.1	-0.3	-0.2	0.0	+1.1*	+0.3
	$(\Sigma v_{ij})^2$	0.04	0.01	0.16	0.16	0.01	0.09	0.04	0.0	1.21	0.09
	$\sum_{i}^{n} (v_{ij}^2)$	0.04	0.07	0.18	0.06	0.01	0.05	0.02	0.0	1.05	0.05
	s^2	0.010	0.022	0.047	0.007	0.002	0.009	0.003	0.0	0.259	0.009
Precision Index = $(0.028/s^2) - 1$		+1.8	+0.3	-0.4	+3.0	+13.0	+2.1	+8.3	∞	-0.9	+2.1

The observed point estimate of bias in performing test Gravity tests on the four products included in the pooled test results is shown in Table XVIII for each of the ten laboratories.

Analyzing the data for accuracy, we proceed to estimating relative accuracy, determine the absolute value of the average algebraic deviation from the mean ($\bar{Y}_j$) for each of the j laboratories, including all data in the computation. The value obtained from this computation will be called the Accuracy Coefficient, A.C.$_j$. Extreme data values are included in the computation because the objective is to estimate how close repeated test results are to the true value of the measured property on the average.

The ASTM reproducibility amount, R.A., can again be used as the basis for establishing a minimum standard for the relative accuracy of test results. Substituting in (4-19) the ninety five per cent confidence limit range for $\bar{Y}_j$ is given by:

$$\bar{Y}_j \leq 0 \pm R.A. \qquad (4-30)$$

Therefore:

$$\text{Minimum Standard for } |\bar{Y}_j| = \frac{R.A.}{2.0} \qquad (4-31)$$

In accuracy Table A.T.$_j$ (1) is to be used for each laboratory as follows:

$$\text{A.T.}_j\text{ (1)} = \text{Minimum Standard for } |\bar{Y}_j| \qquad (4-32)$$

If $A.I._j$ is positive, the laboratory meets the minimum standard established for accuracy. The larger the value of $A.I._j$ the higher the degree of accuracy. If $A.I._j$ is negative, the laboratory does not meet the minimum standard. The larger the negative value is, the more inaccurate are the results obtained by the laboratory.

For the illustrative example, n = 4 and the Reproducibility amount given in the Standard Method of Test for API Gravity of Petroleum Products, ASTM Designation: D 287-55 is 0.5.[24] Substituting in (4-31):

$$\text{Minimum Standard for } \left|\bar{v}_j\right| = \frac{0.5}{2\sqrt{4}} = 0.125$$

and, substituting in (4-32):

$$A.I._j = \frac{0.125}{\left|\bar{v}_j\right|} - 1.0$$

The $\left|\bar{v}_j\right|$ and $A.I._j$ for each of the ten laboratories is computed and tabulated in Table XIII.

Of the ten laboratories, only laboratory 9 with an accuracy index of -0.5 failed to meet the minimum standard for accuracy in the determination of API Gravity of the four products. Of the nine laboratories which are above the minimum standard, laboratories 3 and 4 each with an accuracy index of +0.2 obtained the least accurate measurements while laboratory 8 reported measurements equal to the estimated true API Gravity for all four products.

Analysis of the data for precision. A measure of the variation in the point estimate of the bias is the population variance. Use S_j^2, computed by substitution in (3-18) or its easier computational form (4-27) as the estimator of the variance of measurements made by laboratory j. Include all the data in the computation because the objective is to determine how tightly all the observations reported by the laboratory are clustered. If the objective was to estimate the precision of the test method (as it would be if the standard was being tested) extreme values would be excluded, again pointing out the fact that the proper choice of statistic or estimator is dependent upon what one is trying to measure.

Computation of the variance of measurement, S_j^2, of the API Gravity of the four products of the example is presented in tabular form in Table XIII.

Again using the ASTM Reproducability amount, R.A., as a basis, a minimum standard at the ninety five per cent confidence level can be established for the relative precision of test results.

$$\text{Minimum Standard for } S_j^2 = \left(\frac{R.A.}{3}\right)^2 \qquad (4\text{-}33)$$

A Precision Index, $P.I._j$, can then be computed for each laboratory as follows:

$$P.I._j = \frac{\text{Minimum Standard for } S_j^2}{S_j^2} - 1.0 \qquad (4\text{-}34)$$

If $P.I._j$ is positive, the laboratory meets the minimum standard established for precision. The larger the value of $P.I._j$, the higher the degree of precision. If $P.I._j$ is negative, the laboratory does not meet the minimum standard. The larger the negative value is, the less precise are the results obtained by the laboratory.

For the illustrative example, substituting in (4-33):

$$\text{Minimum Standard for } s_j^2 = \left(\frac{0.5}{3}\right)^2 = 0.028$$

and, substituting in (4-34):

$$P.I._j = \frac{0.028}{s_j^2} - 1.0$$

Computation of the $P.I._j$ for each of the ten laboratories of the example is given in Table XIII.

Two of the ten laboratories, laboratory 3 with a P.I. of -0.4 and laboratory 9 with a P.I. of -0.9, failed to meet the minimum standards for precision in determination of the API Gravity of the four products. Measurements obtained by laboratory 9 were the least precise while those obtained by laboratory 8 were the most precise.

Interpretation of Analysis Results

Accuracy/mistakes. Relative freedom from mistakes is determined by the simple inspection of incidence of extreme values among observations reported by the laboratory. An

excessive number of mistakes indicates possible carelessness. In a laboratory with more than one operator or more than one set of equipment, it may reflect a difference in systematic error among the tests. Since mistakes are due to assignable causes, the established standard for true mistakes should be zero. However, since observations are classified as mistakes on the basis of a statistical decision rule which carries a risk of making a wrong decision, no stigma should accompany infrequent occurrences of "mistakes." For example, a decision rule at the ninety five per cent confidence level will misclassify one chance error out of twenty as a mistake in the long run.

Accuracy/systematic errors. Relatively poor accuracy may be the result of a systematic error or errors. The estimated bias, $\bar{v}_j$, provides a direct measurement of the magnitude and direction of a possible systematic error. A large bias may reflect a local modification to the test method, either intentional, or accidental by reason of misinterpretation. It may also indicate a measuring instrument out of calibration for any reason.

Accuracy/precision. Relatively poor single measurement accuracy may result from relatively poor precision. When relatively poor precision is indicated it may be due to (A) excessive variation in the response of a measuring

instrument, (B) failure to strictly conform with the prescribed test method, or (C) carelessness producing frequent minor mistakes in a random pattern.

Application to the illustrative problem. In the illustrative example, examination of the data indicates a single gross blunder as the probable cause of the failure of laboratory 9 to meet the minimum standard for accuracy. There is no convincing evidence of a significant bias error affecting measurements and three of the four measurements appear free of mistakes.

Laboratory 3 meets the minimum standard for accuracy but not for precision. Poor precision could result in poor accuracy of any single measurement and the laboratory should review the test method to insure that it is being strictly followed.

Laboratory 4 is within limits of both precision and accuracy but shows an apparent bias. Since bias is due to assignable causes, the laboratory should attempt to discover the cause and eliminate it.

LABORATORY RANKING INDEX

Discussion

An index for indicating the relative reliability of a laboratory in the performance of a specified test on a given

product or homogeneous group of products was described in the preceding section. Laboratories can also be rated according to their relative reliability in performance of the family of tests associated with a single product. This would be a useful refinement on the Summary of Laboratory Performance described in Chapter I, in that it would supply a direct performance standard for command personnel in evaluating laboratories under their jurisdiction. To provide the most efficient indication of operational effectiveness to the military commander, consideration should be given to the fact that certain properties of each product have greater significance in regard to the operational performance of the product than other properties. This importance can be recognized by assigning weighting factors to each test.

The measure of relative accuracy common to all tests is the normal deviate, z_{ij}. An appropriate Laboratory Ranking Index, LRI_j, for laboratory j then would be the total of the weighted z_i's computed for each of the n tests.

$$LRI_j = \sum_{i}^{n} w_i \, z_{ij} \qquad (4\text{-}35)$$

Where:

w_i = the weighting factor for test i determined by the relative significance of that test to the operational performance of the product

And:

$$z_{ij} = \frac{x_{ij} - \hat{u}_i}{\hat{\sigma}_i} \qquad (4\text{-}36)$$

The w_i's are arbitrarily chosen as positive and if these factors are normalized, i.e. $\Sigma w_i = 1$, the Laboratory Ranking Index will have the same units as z and will represent a weighted average.

Tests which are not adaptable to inclusion, notably those which require qualitative rather than quantitative observations such as the test for copper strip corrosion by petroleum products,[28] can be excluded from determination of the Laboratory Ranking Index by assigning a weighting factor of zero.

Procedure

Data and assumptions. The raw data required are the results (for a sample of a given product) of all tests, n in number, performed on the product at each of m laboratories. The same assumptions made in preceding sections of this chapter regarding homogeneity of the sub-divided samples, normal frequency distributions of observations and proven test procedures apply.

The procedure for determining the ranking index for each laboratory will be illustrated utilizing correlation test data reported for sample 64-31 of Ashless Dispersant

Aircraft Lubricating Oil. It is arbitrarily assumed that only five tests have been assigned a non-zero weighting factor. These five sets of test results and non-significant weighting factors assigned for illustrative purposes only are listed in Table XIV.

Computing the normal deviate. One estimates the true value of the property for each test. Extreme values resulting from bias errors or mistakes must be excluded from the computation. Test suspected outliers by Dixon's ratio test [equation (3-15) or (3-16)] and use the arithmetic mean as the estimator of μ. As an alternative, the Average of the Best Two estimator of μ, taken from Table II, can be used to facilitate computation.

Suspected extreme values in the illustrative data of Table XIV were tested by Dixon's method and the observation 0.232 submitted by laboratory 1 for test 5 (Carbon Residue) was rejected as significant at the ninety five per cent confidence level. The arithmetic mean estimates of μ_i are shown in the table.

One computes the algebraic deviation, v_{ij}, from the mean of test i and divides by the estimated standard deviation of the population of laboratory test results, $\hat{\sigma}_i$, to determine the normal deviate, z_{ij}. The efficient estimator of the standard deviation computed from equation (3-19) may be used. If it is desired to simplify computation by the

use of one of the less efficient estimators, the Modified Linear estimator given in Table II is recommended.

Values of v_{ij}, σ_i and z_{ij} computed for the illustrative data are shown in Table XIV.

Computing the ranking index. The Laboratory Ranking Index, LRI_j, is computed from equation (4-35). The laboratory with the smallest LRI is the most accurate in the overall measurement of the product's properties.

The LRI's for Aircraft Lubricating Oil computed for the ten laboratories in the example are shown in Table XIV. Laboratory 6, with an LRI of 0.346, ranks best among the ten, while laboratory 1, with an LRI of 1.697, ranks lowest. One interpretation that can be given to this relationship is that the probability that laboratory 6 will properly classify oil on the borderline of acceptability as the result of a single set of tests is considerably higher than that of laboratory 1.

TABLE XIV

COMPUTATION OF LABORATORY RANKING INDEX OF TEN LABORATORIES FOR TESTING OF AIRCRAFT ENGINE LUBRICATING OIL (ASHLESS DISPERSANT)

i Test \ Lab. j	1	2	3	4	5	6	7	8	9	10
1 API Gravity										
X_{1j}	26.7	27.0	26.8	27.0	27.0	26.9	26.9	27.1	27.0	27.2
	$\hat{\mu}_1 = 26.96$									
v_{1j}	-0.26	+0.04	-0.16	+0.04	-0.04	-0.06	+0.06	+0.14	+0.04	+0.24
	$\hat{\sigma}_1 = 0.1574$									
Z_{1j}	1.65	0.25	1.02	0.25	0.25	0.38	0.38	0.89	0.25	0.52
	$w_1 = .2$									
$w_1 Z_{1j}$	.330	.050	.204	.050	.050	.076	.076	.178	.050	.104

TABLE XIV (continued)

i Test \ Lab. j	1	2	3	4	5	6	7	8	9	10
2 Viscosity										
X_{2j}	135.45	134.00	133.53	133.89	127.04	133.24	133.00	133.45	133.31	133.25
	$\hat{\mu}_2 = 133.02$									
v_{2j}	+2.43	+0.98	+0.51	+0.87	-5.98	+0.22	-0.02	+2.43	+0.29	+0.23
	$\hat{\sigma}_2 = 2.137$									
z_{2j}	1.14	0.46	0.24	0.41	2.80	0.10	0.01	1.14	0.14	0.10
	$w_2 = .3$									
$w_2 z_{2j}$	.342	.138	.072	.123	.840	.030	.003	.342	.042	.030
3 Flash Point										
X_{3j}	530	550	525	550	535	540	545	545	550	550
	$\hat{\mu}_3 = 540$									
v_{3j}	-10	+10	-15	+10	-5	0	+5	+5	+10	+10
	$\hat{\sigma}_3 = 6.888$									
z_{3j}	1.45	1.45	2.18	1.45	0.72	0	0.72	0.72	1.45	1.45
	$w_3 = .1$									
$w_3 z_{3j}$	0.145	0.145	0.218	0.145	0.072	0	0.072	0.072	0.145	0.145

TABLE XIV (continued)

i Test \ Lab. j	1	2	3	4	5	6	7	8	9	10
4 Ash										
X_{4j}	0.0024	0.0012	0.0	0.0	0.0019	0.0014	0.0002	0.0016	0.0006	0.0026
$\hat{\mu}_4 = 0.0012$										
v_{4j}	+0.0012	0	-0.0012	-0.0012	+0.0007	+0.0002	-0.0010	+0.0004	-0.0006	+0.0014
$\hat{\sigma}_4 = 0.000984$										
z_{4j}	1.22	0	1.22	1.22	0.71	0.20	1.02	0.41	0.61	1.42
$w_4 = .2$										
$w_4 z_{4j}$	.244	0	.244	.244	.142	.040	.204	.082	.122	.284
5 Carbon Res.										
X_{5j}	0.232*	0.351	0.317	0.362	0.439	0.321	0.370	0.403	0.369	0.327
$\hat{\mu}_5 = 0.362$ (excludes X_{51})										
v_{5j}	-0.130*	-0.011	-0.045	0	+0.077	-0.041	+0.008	+0.041	+0.007	-0.035
$\hat{\sigma}_5 = 0.0409$										
z_{5j}	3.18	0.27	1.10	0	1.88	1.00	0.20	1.00	0.17	0.86
$w_5 = .2$										
$w_5 z_{5j}$	.636	.054	.220	0	.376	.200	.040	.200	.034	.172

TABLE XIV (continued)

i Test \ Lab. j	1	2	3	4	5	6	7	8	9	10
$LRI=\sum_i w_i z_{ij}$	1.697	0.387	0.958	0.562	1.480	0.346	0.395	0.874	0.393	0.735

*Extreme value--significant at the 95 per cent confidence level.

CHAPTER V

ANALYSIS BY A GRAPHICAL METHOD

A graphical method for evaluating new laboratory test procedures has been proposed by Youden.[29] This method utilizes the median as a measure of central tendency. As a measure of variability, it utilizes an unbiased estimate of standard deviation based on the mean difference of paired results. Using this technique as a foundation, a graphical method for evaluating the relative accuracy and precision of a group of testing laboratories utilizing specified, proven test procedures will be developed in this chapter.

Correlation test data will be analyzed by this method to illustrate the potential usefulness to a military commander exercising quality surveillance over a group of widely scattered laboratories.

DISCUSSION

In the target analogy, the reliability problem was defined as one of consistently coming as close as possible to the intersection of the horizontal and vertical hairlines. Assuming the unattainable situation of absence of all error, laboratory test results would invariably be the true value of the property being measured. However, the existence of various sources of error has been acknowledged.

Consequently, even under the best possible circumstances, the measurement obtained is expected, with a given degree of confidence, to be only one of an infinite number of values within a statistically determinable range.

Assume first that errors do exist but that only mistakes or systematic errors are possible; none are due to chance causes. Relating this to the definitions given to precision and accuracy, the assumption is one of perfect precision but possibly poor accuracy. The true value of a property being measured can be represented by either a horizontal or a vertical centerline. An observed value of the property can then be represented by a point at a perpendicular distance from the centerline, which distance measures the inaccuracy of the observation. Such a representation is illustrated in Figure 5-1.

Assume now that two observations are to be made of the same property. The first observation is to be plotted on a horizontal axis and the second is to be plotted on a vertical axis. If the two axes are overlaid, a graph subdivided into four quadrants as shown in Figure 5-2 results. The quadrants have been numbered counterclockwise from I to IV starting with the upper right-hand quadrant in the conventional manner. Let the true value of the property being measured be zero and let the horizontal axis be identified as the A-axis and the vertical axis as the B-axis. Both axes are

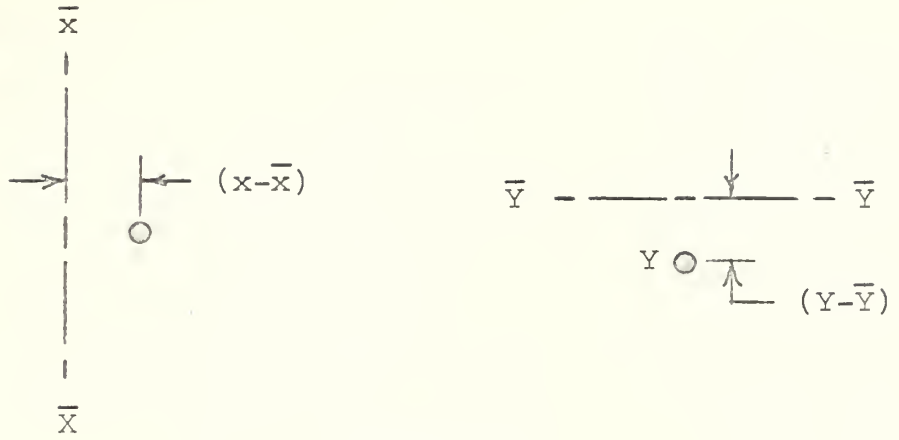

FIGURE 5-1

DEVIATION FROM A HORIZONTAL OR VERTICAL AXIS

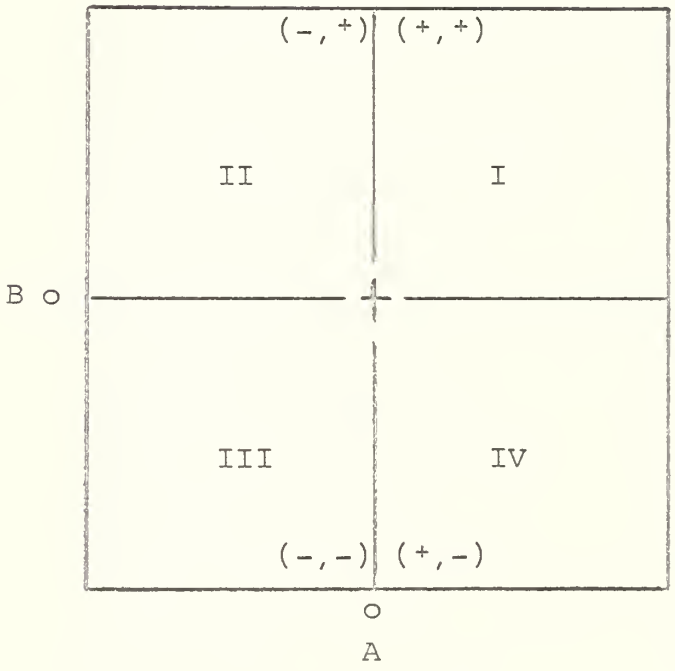

FIGURE 5-2

QUADRANTS FORMED BY THE INTERSECTION
OF TWO PERPENDICULAR AXES

to the same scale. The two observations will be identified as A and B respectively.

Recalling that chance errors are impossible, if no mistakes or systematic errors occur both observations will be the true value, placing data point (A,B) at the intersection of the two axes. The presence of only a systematic error will result in data point (A,B) appearing in either quadrant I if the error causes observations higher than the true value, zero, or in quadrant III if the error causes observations lower than the true value, zero. The appearance of a data point (A,B) in quadrant II or IV results from one observation being greater than and one observation being less than the true value. This can be explained only on the basis of a mistake since systematic errors produce a constant bias and random errors have been disallowed.

Now discount the possibility of mistakes as well as random errors. As a consequence, data points can occur only in quadrant I or III if a systematic error is causing a positive or negative bias respectively, or at the intersection of the axes if there is no systematic error. In fact, since the systematic error has a constant value, the locus of all possible data points is a straight line passing through the intersection of the A and B axes and bisecting quadrants I and III as shown in Figure 5-3.

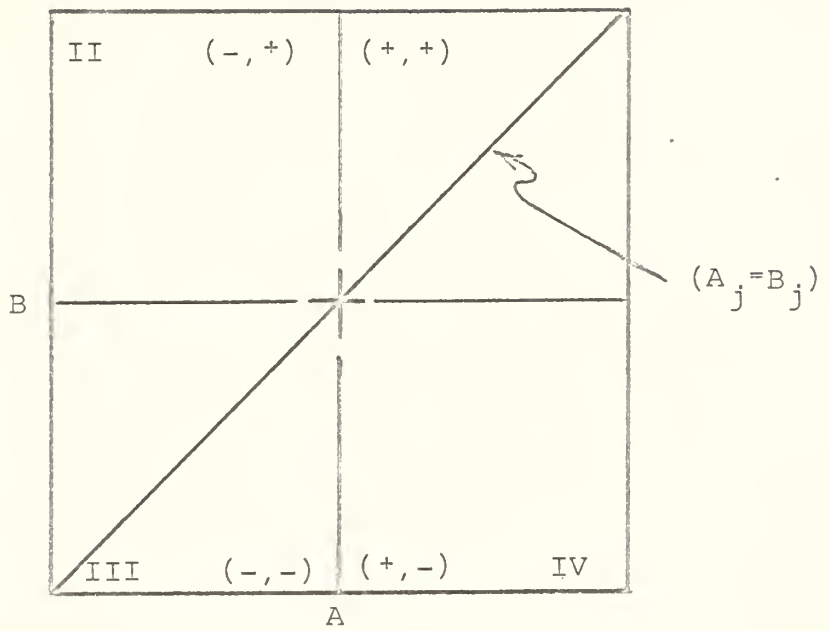

FIGURE 5-3

THE LOCUS OF EXPECTED VALUES FOR ALL OBSERVATIONS (A_j,B_j) AFFECTED ONLY BY SYSTEMATIC ERRORS

The locus is a straight line through the intersections of the A and B axes bisecting quadrants I and III.

As the next step, recognition is given to the existence of chance causes of variation which will cause deviations from the locus just described. Excluding the possibility of mistakes, a data point (A,B) is now expected to fall not on the forty-five degree line through the intersection of the axes but within an area surrounding a given point on the line. The maximum amount by which a pair of observations can be expected to vary a stated percentage of the time solely due to chance causes can be determined and a circle of statistical confidence can be constructed around each point on the line.

The consistent recurrence of scattered paired data points within such a circle centered on the intersection of the two axes would indicate highly reliable performance. The observations would be considered accurate because they are clustered around the true values of A and B. They are acceptably precise because they vary only within the limits of the established performance standard. The consistent recurrence of paired data points within such a circle of confidence centered far out on the forty-five degree line would indicate an acceptable degree of precision but poor accuracy. The accuracy is considered poor because the paired observations are centered on a point far removed from the true values of A and B (Figure 5-4).

Since the forty-five degree line is the locus of an infinite number of points, the circles of confidence around them become a confidence band bounded by parallel lines on each side of the forty-five degree line at a perpendicular distance equal to the radius of the circle of confidence (Figure 5-5).

As a final consideration, assume the existence of a large group of laboratories, each having only one operator and one set of equipment. Also assume once again the existence of only random errors so that all data points will cluster about the intersection of the true value axes. Two variances can then be determined. The repeatability

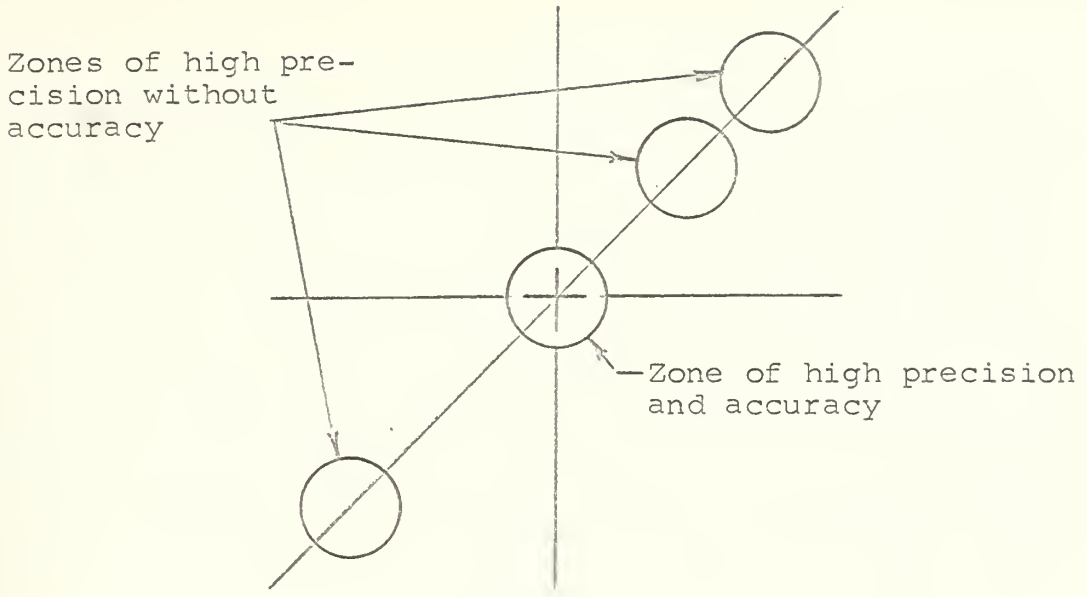

FIGURE 5-4

ZONES OF VARIABILITY ESTABLISHED BY SETTING ARBITRARY STANDARDS FOR MEASURING ACCEPTABLE PRECISION LIMITS

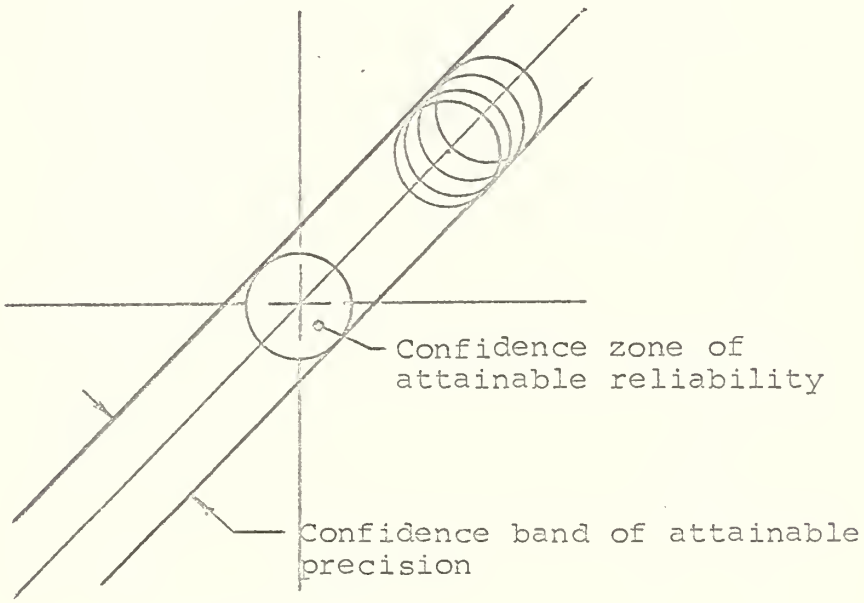

FIGURE 5-5

DEVELOPMENT OF THE CONFIDENCE BAND FOR PRECISION

variance for the test is the random variance between repeat measurements by the same operator using the same equipment in the same laboratory. The reproducibility variance for the test is the variance between measurements obtained at different laboratories. The reproducibility variance will normally be larger than the repeatability variance because of the introduction of additional sources of random variation.

Setting Confidence Limits

The horizontal deviations from the estimated true population value, $\bar{A}$, and the vertical deviations from the estimated true population value, $\bar{B}$, are independent and normally distributed and have a common standard deviation for the population or for any particular laboratory. The probability that a data point (A_j, B_j) is within b standard deviations of the point of intersection of the two axes $(\bar{A},\bar{B})$ can be determined by integration in polar coordinates.[30] The expression which results is:

$$Pr\ (b\sigma) = 1 - \exp\left(\frac{-r^2}{2\sigma^2}\right) \qquad (5\text{-}1)$$

$$= 1 - \exp\left(\frac{-b^2\sigma^2}{2\sigma^2}\right)$$

$$= 1 - \exp\left(\frac{-b^2}{2}\right) \qquad (5\text{-}2)$$

where r = the radial distance to data point (A_j,B_j) = $b\sigma$

By rearranging terms, an expression is obtained for computing the limiting value of b for any desired confidence level.

$$\text{Confidence Level, C.L.} = \Pr(b\sigma) = 1 - \exp\left(\frac{-b^2}{2}\right)$$

$$\exp\left(\frac{-b^2}{2}\right) = 1 - \text{C.L.} \qquad (5\text{-}3)$$

Taking logarithms of both sides:

$$\frac{-b^2}{2} = \ln(1 - \text{C.L.})$$

$$\frac{b}{\sqrt{2}} = \sqrt{-\ln(1 - \text{C.L.})}$$

$$b = 1.414\ \sqrt{-\ln(1 - \text{C.L.})} \qquad (5\text{-}4)$$

The radius of a circle of confidence around the intersection of the two means, $r_{C.L.}$, can also be computed.

$$r_{C.L.} = b\sigma = 1.414\ \sigma\ \sqrt{-\ln(1 - \text{C.L.})} \qquad (5\text{-}5)$$

The radius for a ninety five per cent confidence level is:

$$r_{0.95} = 1.414\ \sigma\ \sqrt{-\ln(1 - 0.95)}$$

$$= 1.414\ \sigma \cdot \sqrt{-(-3)}$$

$$= 2.45\ \sigma \qquad (5\text{-}6)$$

The ninety five per cent level for the difference between two observations is 2.77σ.[31] Using the ASTM Reproducibility amount as a standard, for single observations:

$$R.A._{.} = 2.77\ \sigma_X$$

$$\sigma_X = \frac{R.A._{.}}{2.77} \tag{5-7}$$

For the difference between averages of two pairs of observations (or between the average of two observations and the average of the two means):

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{2}} = \frac{R.A._{.}}{2.77\sqrt{2}} \tag{5-8}$$

Therefore:

$$r_{0.95} = \frac{(2.45)(R.A._{.})}{2.77\sqrt{2}}$$

$$= 0.386\ \frac{(R.A._{.})}{\sqrt{2}}$$

$$= 0.625(R.A._{.}) \tag{5-9}$$

In order to estimate the precision of individual laboratories' test results, a straight line bisecting quadrants I and III is passed through the intersection of the two median lines at an angle of forty five degrees to the axis. Parallel lines can then be constructed on opposite sides of this forty five degree line to form a ninety five per cent confidence interval or band. For convenience, the limits given in ASTM Standards on Petroleum Products and Lubricants are again used to determine the perpendicular distance from the forty five degree line to the boundary of the confidence band. As before, the correction factor of 0.625 must be applied to convert the amount from a range for

a linear normal distribution to a radius for a circular normal distribution. It may be found to be more convenient to locate points on the limit line by measuring the horizontal (or vertical) rather than the perpendicular distance from the forty five degree line. This distance is determined by multiplying the radius by the secant of forty five degrees, 1.414.

The Reproducibility amount rather than the Repeatability amount was chosen as the basis for determination of the ninety-five per cent confidence limits in order to have a minimum standard applicable to all laboratories. The Repeatability amount is the difference which a pair of results obtained by the same operator using the same equipment should not exceed. Quite obviously, such precision is statistically beyond the reach of a large laboratory if paired results were obtained from different combinations of equipment and operator. The Reproducibility limits are the realistic limits in such cases.

PROCEDURE

Data

The raw data required are the test results for a given property obtained from two samples, A and B, of different batches of product which have each been divided and distributed among the participating laboratories. Although

desirable, it is not absolutely necessary that both samples be of the same product. It may be feasible to pair test results of a sample of motor gasoline with test results of a sample of aviation gasoline for example. The objective is to avoid introducing additional sources of variability. Generally, this objective can be accomplished if the test procedures are identical and if the two samples are reasonably close in the magnitude of the property being evaluated.

Assumptions

Analysis of the data is based upon the following assumptions: (A) The sub-divided samples are homogeneous, that is, there is no quality variation of the material distributed to the various participating laboratories, (B) The universe of observations for each laboratory and all laboratories is normally distributed, (C) The test procedure has been proven, that is, it is adequately described to preclude general misinterpretation of the exact procedure to be followed.

Plotting the Data

Select the paired test results to be plotted for a given property and prepare a graph on rectangular coordinate paper. Using the same units and the same scale on both axis, mark an appropriate range on the X axis and Y axis to cover the range of results submitted for sample A and sample B

respectively. Plot the pairs of results reported by the laboratories.

Correlation test observations of Vapor Pressure on sample 63-02 and sample 63-1701 of Combat Automotive Gasoline will be used to illustrate the procedure. These observations are tabulated in Table XV as Test A and Test B respectively. The paired data points are plotted in Figure 5-6.

Estimating Central Tendency

The estimated true value of the property for sample A and sample B can be determined graphically using the median as an estimator. The median is chosen as the estimator because of the relative ease with which it can be constructed in comparison with the mean or Average of the Best Two. The latter estimators both require computation to evaluate the estimate of the population value. The medians can be determined simply by halving the points. The median of A, represented by the symbol $\tilde{A}$, is a vertical line erected perpendicular to the A axis so that the number of data points on either side of the line is equal as illustrated in Figure 5-7. The median of B, represented by the symbol $\tilde{B}$, is erected perpendicular to the B axis in the same manner.

TABLE XV

CORRELATION TEST OBSERVATIONS OF VAPOR PRESSURE OF FOUR SAMPLES OF COMBAT MOTOR GASOLINE

Test	1	2	3	4	5	6	7	8	9	10
A	6.45	6.65	6.4	6.9	6.6	6.9	6.7	6.75	6.5	6.75
B	6.45	6.85	6.70	6.8	6.51	7.1	7.0	6.93	7.0	6.95
C	6.9	6.86	6.7	6.85	6.93	6.88	6.5	6.65	7.3	7.0
D	6.4	6.5	6.6	6.5	6.7	6.7	6.6	6.6	6.4	6.3

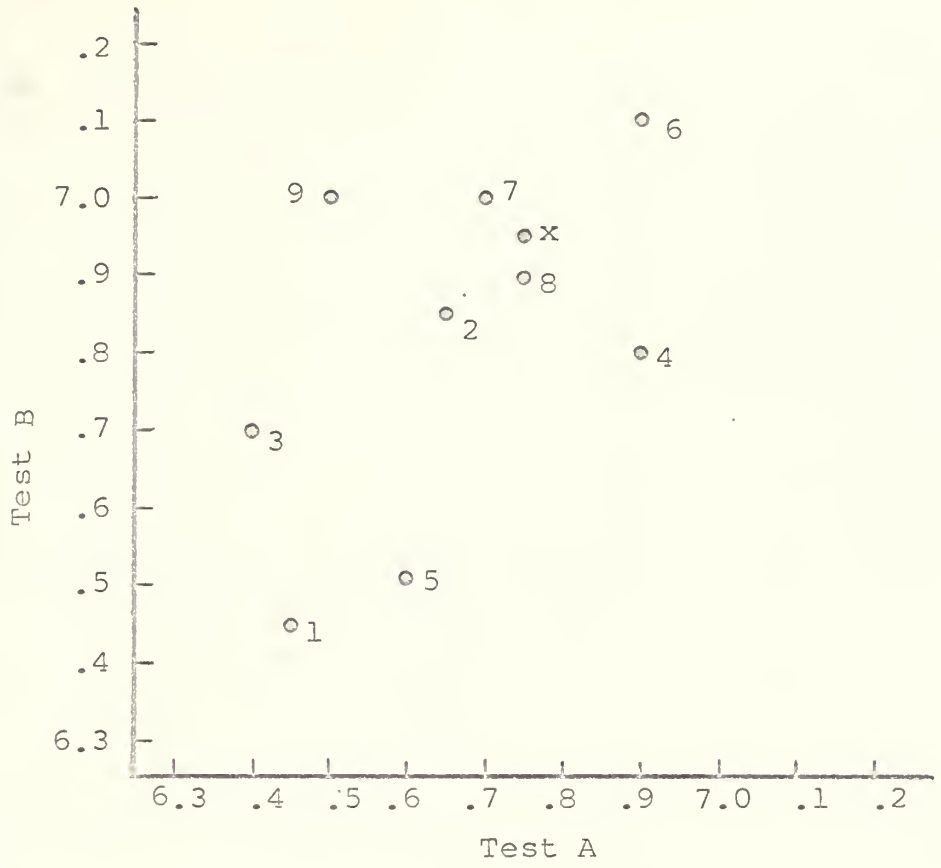

FIGURE 5-6

PLOT OF PAIRED CORRELATION TEST MEASUREMENTS OF VAPOR PRESSURE OF TWO SAMPLES OF COMBAT AUTOMOTIVE GASOLINE

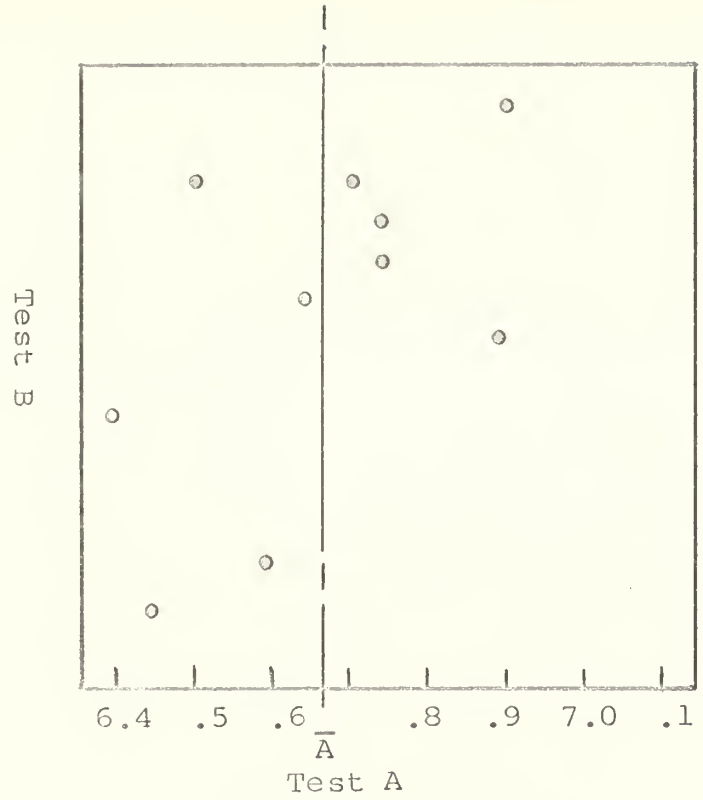

FIGURE 5-7(A)

CONSTRUCTION OF THE MEDIAN LINE OF A VALUES

Fifty per cent of the data points lie on each side of the median, $\bar{A}$.

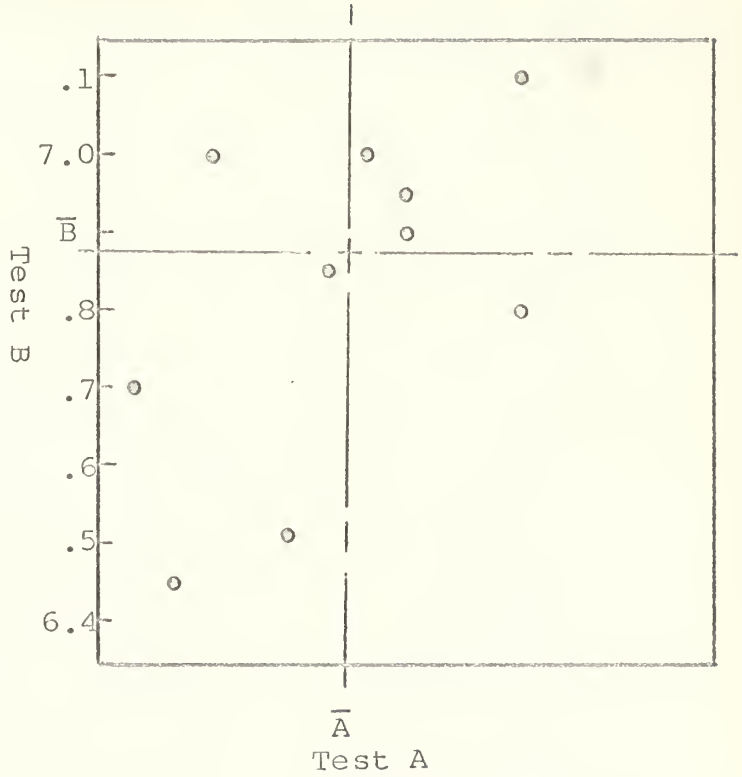

FIGURE 5-7(B)

CONSTRUCTION OF THE MEDIAN LINE OF B VALUES

Fifty per cent of the data points lie on each side of the median, $\bar{B}$.

Setting Confidence Limits

Determine the radius of the ninety five per cent confidence circle, $r_{0.95}$, by substitution in equations (5-6) or (5-9). If equation (5-9) is to be used, determine the Reproducibility amount from the applicable ASTM Standard Method of Test.

The R.A. will be used as the basis for computing $r_{0.95}$ for this example. From the Standard Method of Test for Petroleum Products, ASTM Designation: D323-58,[32] the R.A. for automotive gasoline in the 5 to 16 pound vapor pressure range is 0.3. Substituting in (5-9):

$$r_{0.95} = 0.625\ (0.30)$$
$$= 0.188$$

Construct the ninety five per cent confidence circle for accuracy around the intersection of the median lines $\bar{A}$ and $\bar{B}$, using the radius $r_{0.95}$. With parallel rulers, construct a forty-five degree line (line passing through the intersection of the median lines and bisecting quadrants I and III) and ninety five per cent precision confidence limits parallel to the forty-five degree line and tangent to the ninety five per cent circle for accuracy. Figure 5-8 illustrates the completed graphical construction.

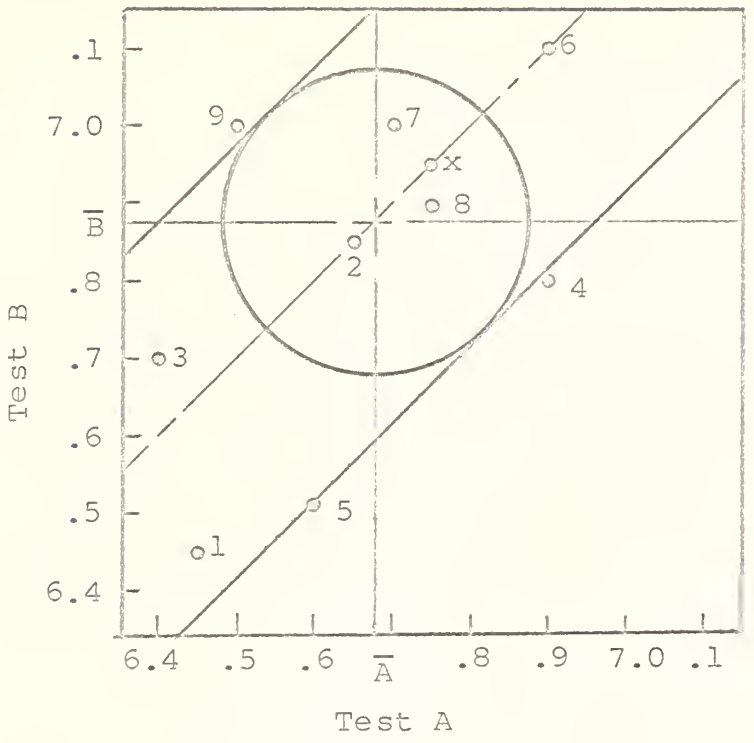

FIGURE 5-8

CONFIDENCE LIMITS FOR ACCURACY AND PRECISION OF DATA PAIRS (A_j, B_j)

The circle is the ninety five per cent confidence limit for accuracy. The parallel lines tangent to the circle are the ninety five per cent confidence limits for precision.

INTERPRETING THE PLOT

Plotted results can be interpreted from either of two viewpoints. The general distribution of data points is of interest in determining the likelihood of sampling errors. The location of individual data points is the basis for laboratory evaluation.

General Distribution of Data Points

If the only errors affecting the data were random errors of precision, positive and negative errors would be relatively small and would occur with equal likelihood. As a result, data points should be expected to be tightly scattered, more or less equally, in all four quadrants formed by the intersection of the two median lines. This is the ideal situation, and is unlikely to occur. Individual laboratory biases will normally cause laboratories to obtain results on the true samples which are either both negative or both positive in relation to the median. A concentration of data points in Quadrant I and Quadrant III can therefore be expected. The more pronounced this tendency to individual bias, the greater the departure will be from the ideal circular distribution.

In the event that the paired observations are nearly equally divided among the four quadrants, the possibility of invalid data resulting from a sample distribution error should be considered.

If the sample divisions distributed to the participating laboratories are not homogeneous as to the property being measured, some will yield high results and some will yield low results. This is true for both samples. The equiprobable set of paired results is:
(high A, high B; high A, low B; low A, high B; low A, low B). It follows that a roughly circular scatter of data points around the intersection of the two medians could be due to heterogeneous divided samples.

Individual Data Points

Data points within the circle surrounding the intersection of the two median lines indicate that the laboratory obtains results for this test which are acceptably accurate, that is, reasonably free from accidental or systematic error. Only five per cent of the time will a pair of observations whose accuracy is affected only by random errors fall outside this circle. Consequently, a data point outside the circle is interpreted as an indication of probable inaccuracy.

Data points within the band surrounding the forty five degree line indicate that the laboratory obtains acceptably precise results for this test, that is, the operators are careful in their work and the results reported are free from careless errors.

Examination of Figure 5-8 shows that the data, used as an example conform to the general distribution pattern normally expected with a tendency to cluster in quadrant I and quadrant III. The dispersion is greater than could be desired however. The indication is that only four of the ten laboratories are measuring the vapor pressure of combat motor gasoline with an acceptable degree of accuracy. Laboratory 2 seems rather precise and accurate, being on the forty five degree line and very close to the intersection of the median lines. The observations reported by laboratory 6 are also highly precise. However the data point appears on the forty five degree line at a considerable distance from the intersection of the median lines and well outside the circle of ninety five per cent confidence for accuracy. It is noted that both measurements were the highest submitted among the ten laboratories for each sample. Interpreted in accordance with the standard for minimum accuracy this indicates that vapor pressure measurements of combat motor gasoline by laboratory 6 are inaccurate. The high degree of precision makes it most probable that the inaccuracy is due to a systematic error and the reviewing command should direct the laboratory to check possible sources of the error and take corrective action. The same general conclusions apply to laboratories 1 and 3. The results reported by laboratories 4, 5 and 9 are inconclusive. Standing alone, one can

only speculate that most probably a mistake has entered into one of the measurements of the pair (the measurement of sample B). In the case of laboratories 4 and 9, the location of the data points could be due to a mistake entering into one of the measurements, chance causes normal to the method (one out of twenty measurements will fall outside the ninety five per cent confidence limits in the long run), or poor precision due to modifications of the test method or due to carelessness. None of these possible causes can be considered most probable without additional data.

ALTERNATE PLOTTING METHODS

Additional analysis of relative performance can be made by comparison of multiple sets of paired observations from each laboratory. These observations can be combined and displayed in various ways. Consider, for example, a subset of four observations, (A,B,C,D) representing the results of the same test on four different samples by the same laboratory. The alphabetical sequence indicates the chronological sequence that the tests were performed. The time interval between tests is one month or more. There are two logical ways in which this set of observations can be formed into subsets of paired data. The first way is to combine the observations in chronological pairs, without duplication, to form the subset (AB,CD). The other way is to combine the observations in chronological pairs, with

duplication, to form the subset (AB,BC,CD). The later alternative has the advantage that it shows the path and therefore, the trend of the data points more readily by providing visual continuity from one point to the next.

The plotting procedure already described provides for plotting the paired observations from two samples, A and B, submitted by m activities. It has the time-saving feature that data are plotted directly as submitted, without preliminary computation and a measure of central tendency, the median, can be determined graphically. Additional pairs of observations obtained from other samples, such as BC and CD, must be plotted separately to use this procedure. If it is desired to plot pairs obtained from more than two samples on a single graph for direct comparison, some manipulation is required to align the axes since the median of each sample will be different. This can be accomplished by overlaying graphs so that their axes coincide and tracing all data points onto one graph. Another method is to transfer data points from one graph to another by measuring their distance from the axes. A third method is to determine the median value of the observations submitted for each sample and code the data by converting the observations to algebraic deviations from the median. The data points can then be plotted directly on a prepared graph with intersecting median lines labeled zero.

Correlation test observations of vapor pressure on sample 64-28 and sample 64-3600 of Combat Automotive Gasoline are tabulated in Table XV as Test C and Test D in addition to the two sets Test A and Test B already analyzed as a pair. The paired data points (C_j, B_j) are plotted in Figure 5-9 and the paired data points (C_j, D_j) are plotted in Figure 5-10. All three of the available graphs, Figures 5-8, 5-9 and 5-10 will now be interpreted as a group. By reference to the interpretation of the graphical analysis of the paired data set (A_j, B_j), one can see how the availability of additional data enhances the utility of the method as a management tool.

The test results reported by laboratory 2 are highly accurate and highly precise. A single measurement of the vapor pressure of an automotive gasoline sample could be accepted with a high degree of confidence as being a very close approximation of the true value. No action is required at the command level.

The measurements reported by laboratory 9 show very poor precision. The pattern of alternating relatively large positive and negative variations from the estimated true vapor pressure of the sample indicates that the poor precision is most probably due to either carelessness or failure to follow strictly the method prescribed for the test. Depending on the possibility that the tests were performed

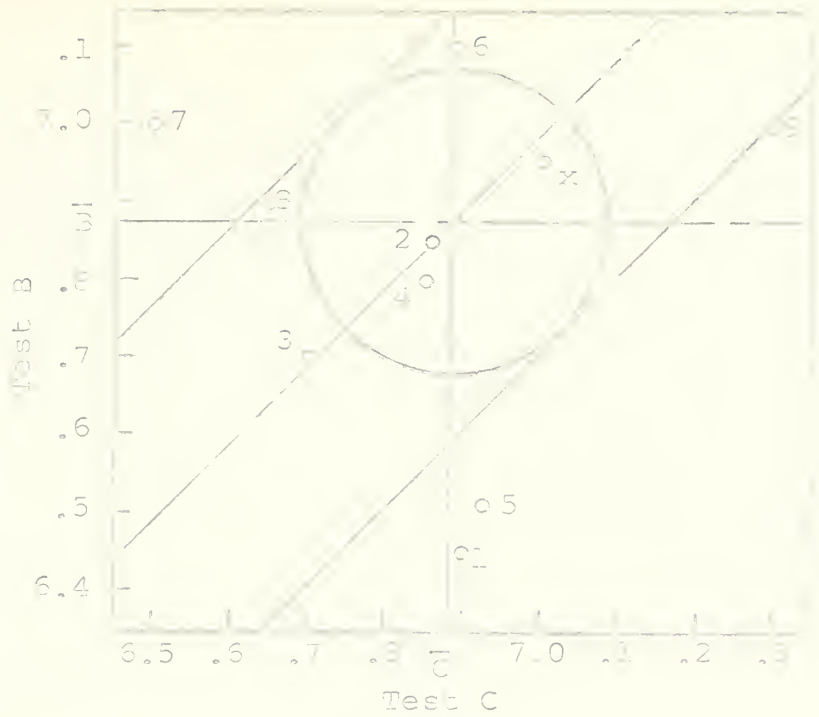

FIGURE 5-9

CONFIDENCE LIMITS FOR ACCURACY AND PRECISION OF DATA PAIRS (C_j, B_j)

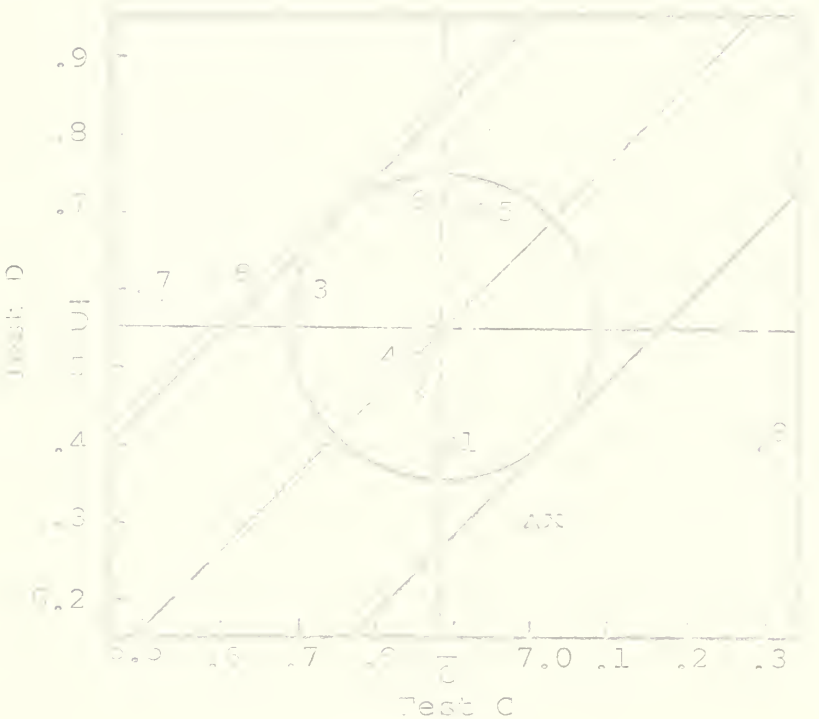

FIGURE 5-10

CONFIDENCE LIMITS FOR ACCURACY AND PRECISION OF DATA PAIRS (C_j, D_j)

by more than one equipment-operator combination (information which would be available to a military commander) a third possibility exists. That is the possibility that the equipment-operator combinations are biased in opposite directions. This possibility is very easily checked from test records. The military commander should direct laboratory 9 to check the precision of its measurements of automotive gasoline by internal investigation and experiment and initiate the necessary action to improve the precision.

Vapor pressure measurements made by laboratory 4 are considered reliable with a high degree of precision and accuracy. Data point (A_4, B_4) was close to the $\bar{B}$ axis although outside the ninety five per cent confidence limits for precision ard accuracy. The indication is that measurement A_4 includes an error due to either a mistake or random causes with about equal probability. No action is required.

Laboratory 5's test results are acceptably accurate and precise. Data point (A_5, B_5) was close to the $\bar{A}$ axis although outside the ninety five per cent confidence limits for precision and accuracy. The indication is that measurements are not quite as precise as those of laboratory 4 but are generally accurate. Judging by its distance from the estimated true vapor pressure, the error of measurement B_4 was most probably due to a mistake but could also have been due to random causes. No action by the military commander is required.

The behavior of the two paired data points of laboratories 7, 8 and 10 is the same as that of laboratories 4 and 5 but in reverse sequence. Laboratories 7, 8 and 10 were within the acceptable limits of precision and accuracy in the earlier period. The latest paired set of measurements from each is outside these limits but lies close to one or the other of the median axes.

Measurement C_6 by laboratory 6 was the same as the $\overline{C}$. The latest measurement, although acceptable as to accuracy by the test, is again the highest measurement submitted for the sample. The indication is that laboratory 6 has not yet located and corrected the source of its systematic error. The military commander should underscore this indication to the laboratory for further attention. The same general interpretation applies to the data reported by laboratories 1 and 3 except that their bias is in the negative direction.

Figure 5-11 has been prepared to show the trend of the data submitted by each activity in regard to accuracy of measurements. These diagrams are constructed to one-half the scale of Figure 5-8, 5-9 and 5-10, and data are posted as deviations from the median to make them compatible. The same interpretations can be derived from this figure as are given above.

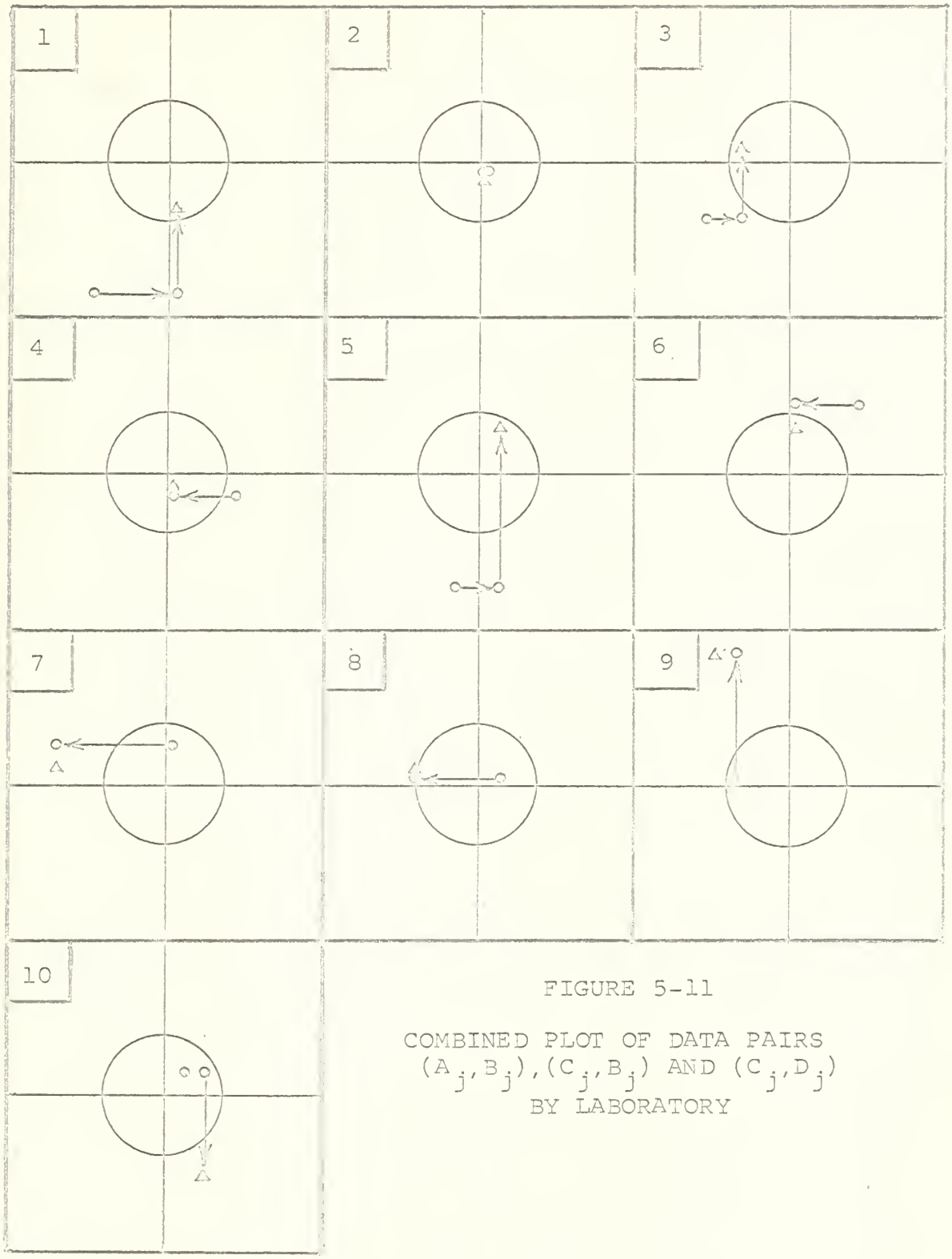

FIGURE 5-11

COMBINED PLOT OF DATA PAIRS
$(A_j, B_j), (C_j, B_j)$ AND (C_j, D_j)
BY LABORATORY

CHAPTER VI

SUMMARY AND CONCLUSIONS

Summary

In this thesis, the author has investigated some statistical means of obtaining more definitive information concerning the reliability of military petroleum testing laboratories than is currently obtained from existing correlation testing programs. Numerical methods of analyzing single observations, paired observations and multiple observations, and a graphical method of analysis were discussed. Procedures were described for analyzing and interpreting the data by each method and were applied to actual military correlation test data.

Table XVI summarizes the tests which can be applied to each activity.

It was found that for a single observation one could test the hypothesis at any pre-selected confidence level that the single observation is statistically the same as the true value of the property being measured. Since there is no dispersion to a single measurement it cannot be tested for precision. Therefore no further amplification can be made of a decision that a single observation is statistically inaccurate at the selected confidence level. This method, using a ninety five per cent confidence level represented

TABLE XVI

SUMMARY OF TESTS OF LABORATORY MEASUREMENTS

Tests based on the ASTM Reproducibility amount (R.A.) provide confidence at the ninety five per cent level.

Tests of Single Observations

Hypothesis test for accuracy:

$$\left(\hat{\mu} - \frac{R.A.}{2}\right) \leq X_{0.95} \leq \left(\hat{\mu} + \frac{R.A.}{2}\right) \qquad (4\text{-}6)$$

Tests of Paired Observations

Hypothesis test for precision:

$$\left|v_{1j} - v_{2j}\right| \leq R.A. \qquad (4\text{-}16)$$

Hypothesis test for accuracy (if precision hypothesis is accepted):

$$-\frac{R.A.}{2\sqrt{2}} \leq \bar{v}_j \leq +\frac{R.A.}{2\sqrt{2}} \qquad (4\text{-}20)$$

Estimate of bias (if precision hypothesis is accepted):

$$\bar{v}_j = \frac{v_{1j} + v_{2j}}{2} \qquad (4\text{-}17)$$

Tests of Multiple Observations

Precision index:

$$P.I._j = \frac{\text{Minimum Standard for } S_j^2}{S_j^2} - 1.0 \qquad (4\text{-}34)$$

$$\text{Minimum Standard for } S_j^2 = \left(\frac{R.A.}{3}\right)^2 \qquad (4\text{-}33)$$

TABLE XVI (continued)

Accuracy index:

$$A.I._j = \frac{\text{Minimum Standard for } \left|\bar{v}_j\right|}{A.C._j} - 1.0 \qquad (4\text{-}32)$$

$$A.C._j = \left|\bar{v}_j\right| \text{ including all } v_j$$

$$\text{Minimum Standard for } \left|\bar{v}_j\right| = \frac{R.A.}{2\sqrt{n}} \qquad (4\text{-}31)$$

Estimate of bias:

Bias estimate = $\bar{v}_j$ excluding extreme values

Laboratory Ranking Index

$$LRI_j = \sum_i^n w_i z_{ij} \qquad (4\text{-}35)$$

w_i = the weighting factor for test i determined by the relative significance of that test to the operational performance of the product.

$$z_{ij} = \frac{X_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \qquad (4\text{-}36)$$

Graphical Analysis

Radius of circle of confidence for accuracy or precision:

$$r_{0.95} = 0.625 \text{ (R.A.)} \qquad (5\text{-}9)$$

by the ASTM Reproducibility amount, is the current method of evaluating correlation test data.

When two homogeneous sets of single observations were pooled and analyzed as pairs of data, the hypothesis that the two single observations of each pair came from the same population could be tested at the ninety five per cent confidence level, thereby measuring the relative precision of the two observations. If the precision hypothesis was accepted, the hypothesis that the average of the two paired observations came from the same population as the estimated $\hat{\mu}$ could be tested to determine the accuracy of the measurements. Again on the prior condition that the precision hypothesis was accepted, the average bias error of the two observations could be determined as an indication of a systematic error due to assignable causes.

When several homogeneous sets of single observations were pooled and analyzed as a group, it was found that precision, accuracy and bias could be measured independently, that is, the validity of the test of one quality of the measurements had no dependence on the prior outcome of another. Precision and accuracy were each measured by an easily interpreted index computed by comparison to established minimum standards. The sign and the magnitude of the index indicates the relative goodness or poorness of accuracy or precision.

A graphical method of analysis was developed which requires only one simple multiplication calculation for its initial application and no mathematical calculations thereafter. The data are analyzed in pairs requiring a minimum of two sets of single observations. When the analysis is limited to two single observations it was found that the same limitations were encountered in interpreting the results when the pair of observations were not adequately precise as were encountered with the numerical analysis of paired observations. Increasing the number of sets of single observations included in the analysis permitted more specific interpretation. When utilizing the graphical method of analysis, the homogeneity of data sets could be verified by observing the general pattern formed by the plotted data. A separate statistical test of homogeneity of variance was required when using the numerical method.

Analysis by the graphical method was used to illustrate how the pooling of homogeneous test data sets increased the effectiveness of analysis of correlation test results as a management tool of the military commander.

In the final analysis, the benefits of reliability in performance of specific tests for specific properties are in correctly classifying a product as to suitability for use. A method of rating laboratories according to their relative reliability in performance of the family of tests

associated with a single product was therefore developed as a useful improvement on the Summary of Laboratory Performance. The method provides for the computation of a Laboratory Ranking Index which is a composite of the relative accuracy of measurement of the various properties of the product, weighted in accordance with their significance in regard to the operational performance of the product.

Conclusions

The current method of analyzing correlation test data is statistically too primitive to provide the military commander with adequate intelligence concerning the effectiveness of the petroleum testing laboratories within his area of jurisdiction.

Maintaining a high degree of accuracy among the petroleum testing laboratories is the specific goal of a military correlation testing program. But accuracy is a function of precision and bias. By analyzing the accuracy of a laboratory's work in terms of precision and bias the correlation testing program can be made into a more effective management-by-exception tool. This requires, as a minimum, analysis of paired homogeneous data sets or, preferably, analysis of multiple homogeneous data sets.

Further investigation of the requirements of an effective correlation testing program is strongly recommended.

This thesis was limited to investigation of some statistical methods of evaluating the reliability of results of laboratory tests of petroleum products and better methods of evaluation were found. Many other facets remain to be explored before a complete program can be formulated and recommended for implementation. Evaluations of optimum frequency of tests, evaluation of the significance of each test, investigation of the validity of using the ASTM Reproducibility amount as a standard, and investigation of the relationship between correlation test measurements and routine test observations are a few.

REFERENCES

[1]J. T. Walter, "How Reliable are Lab Analyses?," Petroleum Refiner, February, 1956, p. 106.

[2]Ibid., p. 107.

[3]Bureau of Naval Personnel, Fundamentals of Petroleum (NAVPERS 10883), Washington: U. S. Government Printing Office, 1953, p. 55.

[4]J. B. Scarborough, Numerical Mathematical Analysis, Baltimore: The Johns Hopkins Press, 1962, p. 378.

[5]A. J. Duncan, Quality Control and Industrial Statistics, Homewood, Illinois: Richard D. Irwin, Inc., 1959, p. 604.

[6]W. J. Dixon, "Ratios Involving Extreme Values," Annals of Mathematical Statistics, March, 1951, pp. 68-78.

[7]W. J. Dixon and F. J. Massey, Introduction to Statistical Analysis, New York: McGraw-Hill Book Co., Inc., 1957, p. 76.

[8]Ibid.

[9]Ibid., p. 404.

[10]J. M. Juran, Quality Control Handbook, New York: McGraw-Hill Book Company, Inc., 1962, p. 13-13.

[11]E. S. Pearson, "The Probability Integral of the Range in Samples of n Observations from a Normal Population," Biometrika, Vol. 32, 1942, p. 301.

[12]Dixon and Massey, op. cit., pp. 405, 406.

[13]Ibid., p. 74.

[14]Ibid., p. 406.

[15]Ibid., p. 265.

[16]Ibid., pp. 404-406.

[17]Duncan, op. cit., p. 51.

[18]D. J. Cowden, Statistical Methods in Quality Control, Englewood Cliffs, N. J.: Prentice Hall, Inc., 1957, p. 10.

[19]W. Volk, Applied Statistics for Engineers, New York: McGraw-Hill Book Company, Inc., 1958, p. 68.

[20]Duncan, op. cit., p. 51.

[21]Ibid., p. 46.

[22]E. Kurnow, G. J. Glasser, and F. R. Ottman, Statistics for Business Decisions, Homewood, Illinois: Richard D. Irwin, Inc., 1959, p. 75.

[23]Volk, op. cit., p. 113.

[24]ASTM Standards for Petroleum Products and Lubricants, Philadelphia: American Society for Testing and Materials, 1961, p. 152.

[25]Volk, op. cit., p. 111.

[26]W. J. Youden, "Graphical Diagnosis of Interlaboratory Test Results," Industrial Quality Control, May, 1959, p. 25.

[27]ASTM Standards, op. cit., p. 14ff.

[28]ASTM Standards, op. cit., p. 75ff.

[29]W. J. Youden, op. cit., pp. 24-28.

[30]E. L. Crow, F. A. Davis, and M. W. Maxfield, Statistics Manual, China Lake, California: U. S. Ordnance Test Station, 1955, p. 28.

[31]Duncan, op. cit., p. 872.

[32]ASTM Standards, op. cit., p. 166.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anderson, R. L. and Bancroft, T. A. Statistical Theory in Research. New York: McGraw-Hill Book Co., Inc., 1951.

ASTM Manual for Conducting an Interlaboratory Study of a Test Method. Philadelphia, Pa.: American Society for Testing and Materials, 1962.

ASTM Manual on Presentation of Data. Philadelphia, Pa.: American Society for Testing and Materials, 1941.

ASTM Manual on Quality Control of Materials. Philadelphia, Pa.: American Society for Testing and Materials, 1951 (Special Technical Publication 15-C).

ASTM Standards on Petroleum Products and Lubricants. Philadelphia: American Society for Testing and Materials, 1961 (issued annually).

Bingham, R. S. "A Guide to the Use of Statistics in the Chemical Industry," Industrial Quality Control, 15:14-18, September, 1958.

Brownlee, K. A. Statistical Theory and Methodology in Science and Engineering. New York: John Wiley and Sons, Inc., 1960.

Bureau of Naval Personnel, Fundamentals of Petroleum (NAVPERS 10883). Washington: U. S. Government Printing Office, 1953.

Burr, I. W. Engineering Statistics and Quality Control. New York: McGraw-Hill Book Company, Inc., 1953.

Cowden, D. J. Statistical Methods in Quality Control. Englewood Cliffs, N. J.: Prentice Hall, Inc., 1957.

Crow, E. L., Davis, F. A. and Maxfield, M. W. Statistics Manual. China Lake, California: U. S. Naval Ordnance Test Station, 1955.

Dixon, W. J. "Ratios Involving Extreme Values," Annals of Mathematical Statistics, 22:68-78, March, 1951.

Dixon, W. J., and Massey, F. J. Introduction to Statistical Analysis. New York: McGraw-Hill Book Co., Inc., 1957.

Duncan, A. J. Quality Control and Industrial Statistics. Homewood, Ill.: Richard D. Irwin, Inc., 1959.

Feigenbaum, A. V. Total Quality Control. New York: McGraw-Hill Book Co., Inc., 1961.

Grant, E. L. Statistical Quality Control. New York: McGraw-Hill Book Co., Inc., 1952.

Juran, J. M. Quality Control Handbook. New York: McGraw-Hill Book Company, Inc., 1962.

Kurnow, E., Glasser, G. J., and Ottman, F. R. Statistics for Business Decisions. Homewood, Illinois: Richard D. Irwin, Inc., 1959.

McElrath, G. W. and Bearman, J. E. "Some Economic Considerations of Inefficient Statistics," Industrial Quality Control, 15:10-14, September, 1959.

Pearson, E. S. "The Probability Integral of the Range in Samples of n Observations from a Normal Population," Biometrika, 32:301, 1942.

Richmond, S. B. Statistical Analysis. New York: The Ronald Press Company, 1964.

Scarborough, J. B. Numerical Mathematical Analysis. Baltimore: The Johns Hopkins Press, 1962.

Schrock, E. M. Quality Control and Statistical Methods. New York: Reinhold Publishing Corp., 1957.

Simon, L. E. An Engineers' Manual of Statistical Methods. New York: John Wiley and Sons, Inc., 1941.

Snedecor, G. W. Statistical Methods. Ames, Iowa: The Iowa State University Press, 1957.

Symposium on Application of Statistics. Philadelphia, Pa.: American Society for Testing and Materials, 1949 (Special Technical Publication No. 103).

Volk, W. Applied Statistics for Engineers. New York: McGraw-Hill Book Company, Inc., 1958.

Walter, J. T. "How Reliable are Lab Analyses?". Petroleum Refiner, 35:106-108, February, 1956.

Youden, W. J. "Graphical Diagnosis of Interlaboratory Test Results," Industrial Quality Control, 15:24-28, May, 1959.

_____. Statistical Methods for Chemists. New York: John Wiley & Sons, 1951.